Speech Emotion Recognition With Recurrent Neural Networks

Bledea Mihaela Alexandra

1. Introduction

Automatic Speech Recognition (ASR) services can be found everywhere. Voice input interfaces are already used in many applications and personal assistants like Apple Siri, Amazon Alexa, Microsoft Cortana. Nowdays machines can successfully recognize human speech, but when it comes to recognition of emotions from the speech, things get complicated. Robots capable of understanding emotions could provide appropriate emotional responses.

Speech Emotion Recognition (SER) is a powerful task in understanding the characteristics of speech in media. In recent years, the interest for solving the SER problem had a radical growth, but, despite the great progress made in artificial intelligence, we are still far from being able to naturally interact with machines, because machines do not understand emotion states.

1.1. Problem

The main problem that appears, when trying to recognize human emotions is that all humans are different and all emotions are experienced and shown to the outside world in a different manner. Emotions are subjective. During a conversation, humans receive lots of meta-information apart from text. They know the person who is speaking, their intonation, loudness, shades. All these factors might considerably influence the true meaning of a phrase and the emotion behind it. Humans take all these elements into consideration when it comes to process a phrase inside their brain and come with the resulted emotion. There also appears the complexity and cost of a database collection. The recordings are relatively easy to collect. One can take dialogues from films, news, blogs. But in most of the cases, the emotions in these are not genuine and they don't come naturally.

1.2. Motivation

Speech Emotion Recognition can be found in applications for healthcare, gaming experience improvement, marketing, stress monitoring, customer satisfaction, social media analysis and many more. This technology, in the future, may provide people with a device with a personal assistant which is not only able to understand the emotions of its owner, but to respond to them with the same empathy and compassion.

1.3. Objectives

The goal is to create a powerful emotion classifier,which later, will be integrated into a personal assistant application, which not only will be able to recognize the emotions of speaker and respond to them adequately, but also help the speaker in person understand and recognize its emotions. Humans that are aware of their emotions build better relationships, avoid or resolve conflicts better, move past difficult feelings more easily and tend to access more joy.

2. Literature review

Through the years, many different approaches were used in order to solve the SER problem.

One approach uses a Convolutional Neural Network (CNN), which extracts features from the raw signal. The model they used was trained end-to-end and stacked on the top of the CNN a 2-layer Long Short-Term Memory. In their study, they used a different objective function. They exchanged the Mean Squared Error as a loss function with a loss function based on the concordance correlation coefficient, exchange which provided better results. The proposed convolutional recurrent neural network for speech emotion recognition, is nothing but a CNN used to extract features from the raw signals before feeding them to a 2-layer LSTM network from the final prediction. (LSTM) [5]

A second approach uses Deep Neural Networks (DNNs) in order to extract high level features from raw data. In the study they show the extracted features are effective for SER challenge. An emotion state probability distribution is produced for each speech segment using DNNs. The utterance-level features are constructed using these distributions. Here, an extreme learning machine (ELM) is used in order to identify utterance-level emotions. [2]

It is shown that using a deep recurrent neural network, it can be learned both the short-time frame-level acoustic features that are emotionally relevant, as well as an appropriate temporal aggregation of those features into a compact utterance-level representation. These being said, another study proposes a novel strategy for feature pooling

over time. It uses a local attention in order to focus on specific regions of a speech signal that are move emotionally salient. [4] The following study evaluated the proposed solution on the IEMOCAP dataset and it has provided more accurate predictions compared to existing emotion recognition algorithms. This study obtained an accuracy between 52.7% and 63.5% for the different RNN architectures they used.

3. Data

All experiments are carried out with audio recordings from the Toronto Emotional Speech Set (TESS) database.

3.1. Database structure

The TESS Database is a collection of audio clips of 2 women expressing 7 different emotions (happiness, fear, disgust, pleasant surprise, neutral, angry and sadness).

What is really interesting about this data set is the fact it is female only and is of very high audio quality. Most of the other datasets go in the opposite direction by being prevailed by male speakers, fact that leads to imbalance representation. Thus, this data set is very suitable for the emotion classifier which is about to be build.

The data set targeted 200 words. Those were carried in the phrase: "Say the word _" by two actresses, one of age 26 and the other one of age 64. Each phrase was said to portray each of the 7 emotions. There are a total of 2800 audio files.

3.2. Data pre-processing

All audio files from the data set were parsed, creating two lists, one containing the path of the audio files and the other one containing the emotion label of the files.

The acoustic features typically utilized in speech emotion recognition are time-domain features, frequency-domain features, statistical features, deep features and hybrid features. In general, feature extraction depends on the frame involved.

The features extracted and used in this study are: MFCC (Mel-Frequency Cepstral Coefficients), ZCR (Zero-Crossing Rate), Mel Spectrogram and RMS (Root Mean Square Value).

MFCC is a frequency domain feature which is frequently used in Speech Emotion Recognition. It captures the timbral and textural aspects of sound. The advantage they have over spectrograms is that they approximate the human auditory system.

"The Zero-Crossing Rate (ZCR) of an audio frame is the rate of sign-changes of the signal during the frame. In other words, it is the number of times the signal changes value, from positive to negative and vice versa, divided by the length of the frame." [1] ZCR can be interpreted as a measure of the noisiness of a signal.

A spectrogram is a representation of the audio data that gives us information about the magnitude as a function of frequency and time. The Mel Spectrogram is used to provide our models with sound information close to what a human would perceive.

Root Mean Square Value is a metering tool that measures the average loudness of an audio track. It represents the energy of the signal and gives a more accurate look at the perceived loudness of a sound.

All the extracted features are combined together. The final output of the pre-processing step is a 1-dimensional vector containing 170 values.

4. Proposed solution

In this paper a Long Short Term Memory Network (LSTM) is used, in order to solve the Speech Emotion Recognition Problem. LSTMs are a special kind of Recurrent Neural Networks (RNN)

Recurrent Neural Networks are networks that have loops in them, which allow the information to persist. A RNN can be thought of as multiple copies of the same network, each passing a message to a successor.

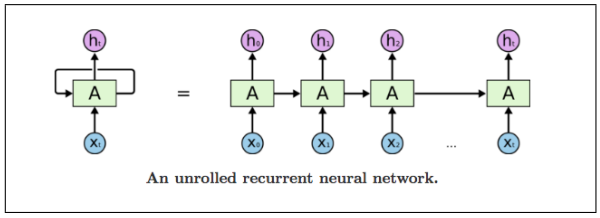


Figure 1. Recurrent Neural Network.

But, RNN's have some limitations. They are thought as networks which can remember at every step, all the information from the previous one. When we want to make some connections between information learned at different times, the gap between can become very large. In theory, RNN's are are absolutely capable of handling such "long-term dependencies", but in practice they don't seem to be able to learn them.

Long Short Term Memory networks (LSTMs) are capable of learning long-term dependencies. They were explicitly designed to avoid this problem, by adding an internal state to the original RNN node. The key of LSTMs is the cell state. This cell has 3 parts, each is a gate. It has the Forget Gate which says what information that was stored in the internal state can be forgotten because is no longer relevant, the Input Gate which says what information should

be added or updated into the storage of the working internal state and Output State which says, out of all the information that is stored in that state, which part of it should be outputted in this particular instance.

Going back to the initial problem we were trying to solve, when it comes to recognition of emotions from audio, studying one sentence at a time is not always sufficient to determine the correct emotion. In most of the cases, all the sentences said at a time should be evaluated and based on the results, the emotion should be predicted. Those being said, this is the reason for choosing to implement an LSTM network.

4.1. Proposed model

The proposed model is a stacked Long Short Term Memory Network and has the following layout:

- Long Short Term Memory layer of 128 units that outputs an array of hidden states for the subsequent layer
- Long Short Term Memory layer of 128 units
- Fully Connected layer of 32 units and Rectified Linear Unit (ReLU) as activation function
- Dropout layer of 0.2 chance
- Fully Connected Final layer of output size 7 and SoftMax as activation function

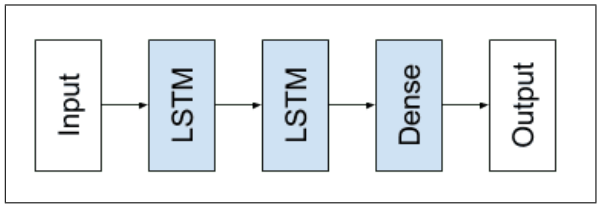


Figure 2. Stacked Long Short-Term Memory Architecture

The output is of size 7, because we've decided to look only into 7 emotions, which are: Sad, Angry, Happy, Surprised, Fear, Disgust and Neutral.

Stacking LSTM hidden layers makes the model deeper, more accurately earning the description as a deep learning technique. Generally, the success of an approach on a wide range of challenging predictions is given to the depth of the neural network.

"The success of deep neural networks is commonly attributed to the hierarchy that is introduced due to the several layers. Each layer processes some part of the task we wish to solve, and passes it on to the next. In this sense, the DNN can be seen as a processing pipeline, in which each layer solves a part of the task before passing it on to the next, until finally the last layer provides the output." [3]

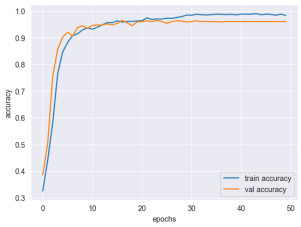
5. Experiments

For all the training and testing scenarios the optimizer, loss function, number of epochs, training set, testing set and the batch size are the same:

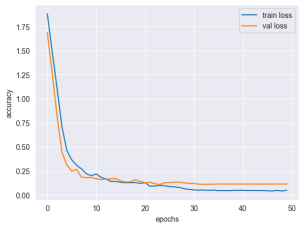
- Optimizer: Adam Optimizer;
- Loss function: Categorical Cross-Entropy. Like that, the RNN will be trained to output a probability over the 7 classes for each audio;
- Learning rate: 0.0010;
- Number of epochs: 50;
- Training set: 2100 utterances;
- Testing set: 700 utterances;
- Batch size: 256;

First Training Scenario

- Extracted features: Mel-Spectrograms were extracted using librosa. The extraction resulted, for each audio file (2800 in total), a vector with 128 values. For the extracted values, the arithmetic mean according to 0 axis was computed;



(a) Accuracy Test 1



(b) Loss Test 1

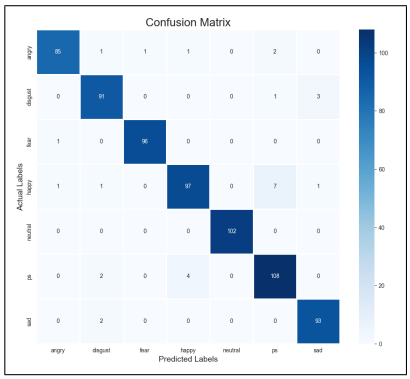
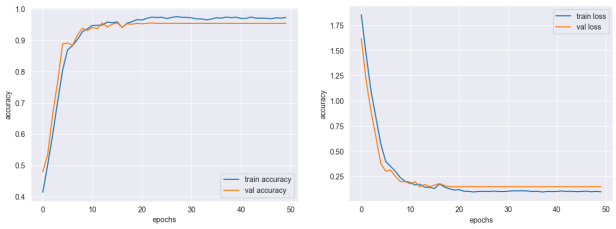


Figure 4. Confusion Matrix Test 1

From graphs we can observe, at the beginning, the accuracy was higher and the loss was lower for the test set compared to the training one. In the end, this situation changed and the accuracy became close to a constant value. The overall accuracy obtain on the test data is 95%. The model is more accurate in predicting neutral, angry and fear emotions.

Second Training Scenario

- Extracted features: Mel-Spectrograms, Mel-Frequency Cepstral Coefficients, Zero-Crossing Rate and Root Mean Square Value were extracted using librosa. The extraction resulted, for each audio file (2800 in total), a vector with 170 values.



(a) Accuracy Test 2 (b) Loss Test 2

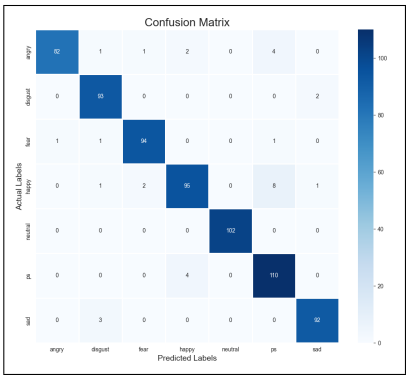


Figure 6. Confusion Matrix Test 2

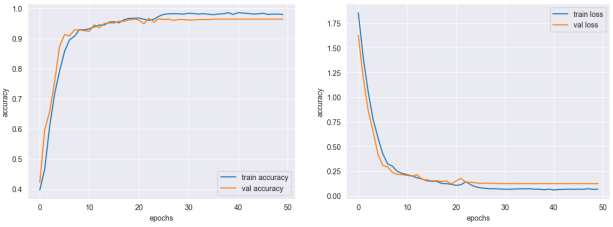
As in the case of first training scenario, we can observe from the graph, the same behavior for the accuracy and loss values.

The overall accuracy obtained on the test data is 96%. The model is more accurate in predicting neutral, angry and fear emotions, just as in the first training scenario. Compared to the first training scenario, using multiple extracted features increased the overall accuracy by 1%.

Third Training Scenario

In this training scenario, we removed the Zero-Crossing Rate from the used features. Since that feature represents the noisiness of a sound and the data set is very clear, we wanted to see if it brought any benefits by extracting and using it.

- Extracted features: Mel-Spectrograms, Mel-Frequency Cepstral Coefficients and Root Mean Square Value were extracted using librosa. The extraction resulted, for each audio file (2800 in total), a vector with 169 values.



(a) Accuracy Test 3 (b) Loss Test 3

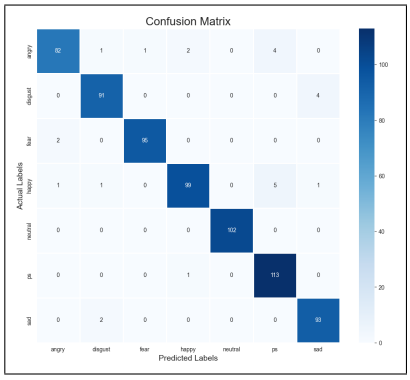


Figure 8. Confusion Matrix Test 3

The overall accuracy obtained on the test data is 96%. In this training scenario, the model is more accurate in predicting neutral, fear and happy emotion, compared to the previous scenario where it was more accurate in predicting angry emotion.

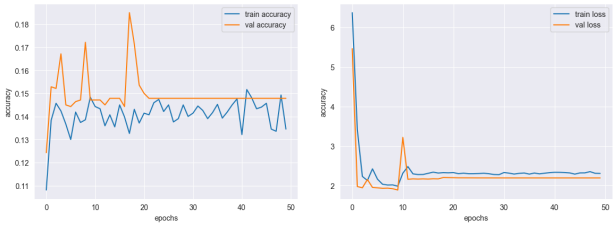
From this result, we can conclude that, for the data set used, the Zero-Crossing Rate is not necessary, because the sound is clear enough and by using it, we don't bring any benefits, just spend more time on feature computation.

Fourth Training Scenario

In this scenario, data augmentation was introduced. In order to check whether the model would work properly on data which is not necessarily clear, but noisy, we injected noise on the dataset and combined it with the original one.

- Extracted features: Mel-Spectrograms, Mel-Frequency Cepstral Coefficients and Root Mean

Square Value were extracted using librosa. The extraction resulted, for each audio file (2800 in total), a vector with 169 values.



(a) Accuracy Test 4 (b) Loss Test 4

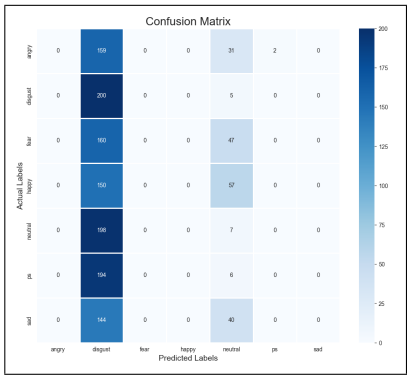


Figure 10. Confusion Matrix Test 4

As we can see from the graphics, in this scenario, we didn't have the same success as in the others. Validation accuracy becomes constant at some point, and is very close to the accuracy for the training set.

The overall accuracy is somewhere between 14% and 16%, value which is extremely small. The model, predicted for most utterances, disgust emotion.

From this experiment, we found a limitation for this approach and created model. It works impressive when the sound is clear, without noise or any distortions, but when it comes to real world sounds, it gives really bad results.

The overall accuracy obtained from the experiments without data augmentation, outperform the results obtained by Kun Han, Dong Yu and Ivan Tashev, in their study [2], where the obtained accuracy is between 50% and 60% and the results obtained obtained by Seyedmahdad Mirsamadi, Emaf Barsoum and Cha Zhang, in their study [4], where the obtained accuracy is between 52.7% and 63.5%.

6. Conclusions

In this paper, we mentioned and compared 3 other studies, which tried to obtain the same thing as we did, ran a series of tests and managed to outperform their results.

In the future, we are looking into the possibility of improving the model, so it could work with data which is not clear and more appropriate to daily voice recordings people do. To do so, the future tests will imply effects such as reverb, distorsion, noise and EQing.

References

[1] Theodoros Giannakopoulos and Aggelos Pikrakis. *Introduction to audio analysis: a MATLAB® approach*. Academic Press, 2014. 2

[2] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*, 2014. 1, 5

[3] Michiel Hermans and Benjamin Schrauwen. Training and analysing deep recurrent neural networks. *Advances in neural information processing systems*, 26, 2013. 3

[4] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2227–2231. IEEE, 2017. 2, 5

[5] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309, 2017. 1