Speech-based emotion recognition
Beszéd alapú érzelemfelismerés

Gergő Emmert, Dóra Gera , András Jankó

Budapest University of Technology and Economics
H-1111 Budapest, Műegyetem rkp. 3-9.

emmertgergo@outlook.com

geradora99@gmail.com
dandrid2@gmail.com

***Abstract*— In this paper, a speech based emotion classification method is presented. Speech emotion recognition is a challenging task since it is unclear what kind of features are able to reflect the characteristics of human emotion from speech. However, traditional feature extractions perform inconsistently for different emotion recognition tasks. Obviously, different spectrogram provides information reflecting difference emotion. This paper proposes an approach to implement an effectively emotion recognition system based on deep convolution neural networks using labeled training audio data. The used database is Ravdess. Eight basic human emotions are investigated including neutral, calm, happy, sad, angry, fearful, disgust and surprised.**

***Összefoglalás*— Ebben a cikkben ismertetjük az általunk használt hangalapú érzelemfelismerés technikákat. A hangalapú érzelemfelismerés nagyon sok kihívást jelent, mert nem egyértelmű, hogy pontosan melyik tulajdonságok tükrözik a beszédben az emberi érzelmeket. A hagyományos jellemző kiemelés technikák nem mindig működnek a különböző fajta érzelemfelismerés problémákra. Az nyilvánvaló, hogy a hang spektrogrammjában benne van a szükséges információ az érzelem megkülönböztetéséhez. Ebben a cikkben bemutatjuk az általunk használt konvolúciós neurális hálózatokat és az általuk elért eredményt. A Ravdess adatbázist használtuk, ami színészek által mondott mondatokat tartalmaz 8 érzelem alapján felcímkézve. A 8 érzelem: semleges, nyugodt, boldog, szomorú, mérges, ijedt, undor, meglepett.**

***Keywords*— speech recognition, emotion recognition, deep learning, keras, tensorflow, python**

I. Introduction

Speech emotion recognition is a challenging task since it is unclear what kind of features are able to reflect the characteristics of human emotion from speech. However, traditional feature extractions perform inconsistently for different emotion recognition tasks. Obviously, different spectrogram provides information reflecting difference emotion. This paper proposes an approach to implement an effectively emotion recognition system based on deep convolution neural networks using labeled training audio data.

II. Description of topic, previous solutions

Reza Chu[1] used a similar approach for speech based emotion recognition. He used MFCC feature extraction to prepare the data and used Convolutional Neural Networks to classify the emotions. He achieved around 70% accuracy.

Dipam Vasani[2] created spectrograms from the audio files, and used Convolutional Neural Networks for the classifications. We reused the spectrogram based solution in our solution too.

III. System design

We have decided to try out multiple solutions to solve the problem, so we can compare the results at the end and it was also a good practice to try out different models.

A. MFCC version

In this version, we have calculated the Mel Frequency Cepstral Coefficient (MFCC) of the audio files and used it as an input for the neural network. We decided to use a self made 1D convolution model. We included dropout and max-pooling in the model. The top of the model used fully connected layers. We used relu activation functions in the model, except for the output, which used softmax activation function. The number of parameters in the model was between 5000 and 250

000. The Final solution included 211 075 parameters.

B. *RESTNET version*

In this version we used transfer learning, and decided to try out the ResNet50 pretrained model, because it's designed to image classification problems. The top layers were added by us, GlobalMaxPooling2D, and 2 Dense layers, the second one is with softmax activation.

C. *Own CNN version*

In this version, we built our own CNN network, assuming that pre-trained nets like ResNet50 were train in quite different images, not spectrograms.

The network consisting of 2-dimensional convolutional layers, max-pooling layers, dropout layers, and dense layers.

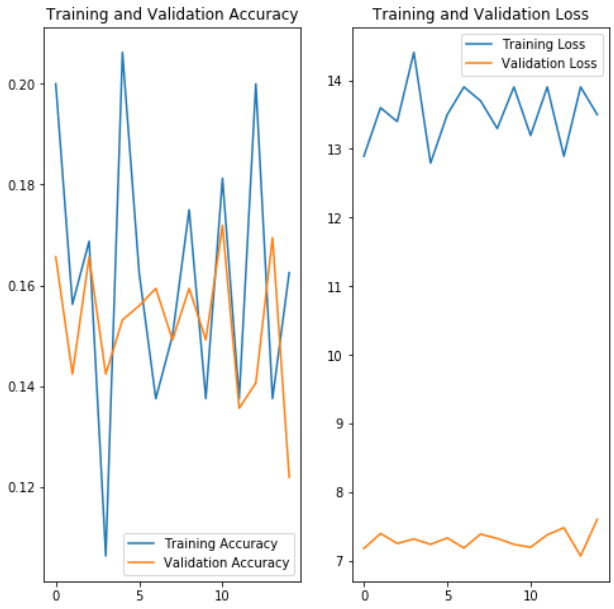
IV. Implementation - RESNET version

A. *Data preparation*

We generated spectrograms from all sound files, and saved them as an image proportional to the lengths. The length of a recording can be significant to recognize an emotion, however a network can only take fixed sized variables. The solution was to slice the images to fix sizes with overlapping . The size of an image is 128x200px.

B. *Learning*

ResNet50 is a convolutional network trained on the ImageNet database. The network was loaded without the top layers, and we used transfer learning. The following image shows the accuracy and loss during the training and the validation process. As the image shows, the model did not really learn the emotions.

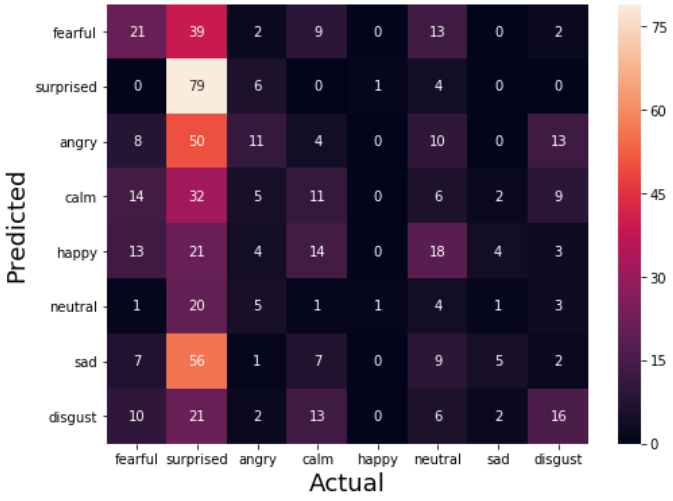


C. *Evaluation*

The best model's training accuracy was 23.646% the loss 1.986, and the validation accuracy was 25.528% and the loss 1.973. During the training we optimized for accuracy and we used rmsprop.

D. *Test*

Created a confusion matrix to show the accuracy in different classes. The matrix shows that the best accuracy was the surprised and fearful emotions. The rest of the emotions were inaccurate.



V. Implementation - MFCC version

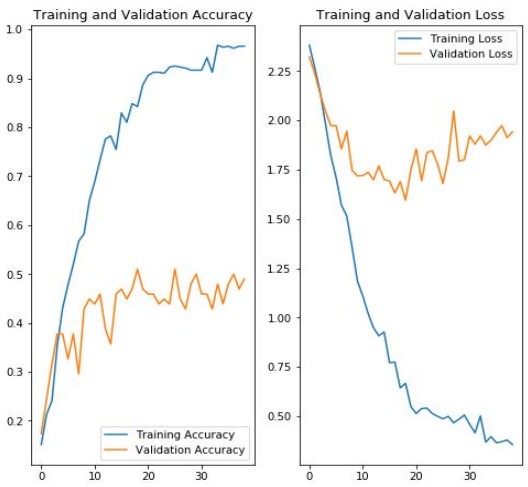
A. Data preparation

During the data preparation, the silence at the beginning and at the end of the audio file was removed. We have also implemented a data augmentation, which added static noise to the audio files.

We have used MFCC feature extraction in this model to help the training. To get the best result, we have excluded the neutral emotion, and only kept the emotions which were labeled as strong intensity.

B. Learning

The following image shows the accuracy and loss during the learning for the training and validation data. As it can be seen, the model unfortunately overfitted the training data. We have tried many ways to prevent this, by using regularizations e.g. dropout, L2 regularization, early stopping, data augmentation and changing the complexity of them model, but they did not help. Probably the problem was that there were not enough training data to produce a better result. During the learning we have optimized for the validation loss instead of the accuracy.



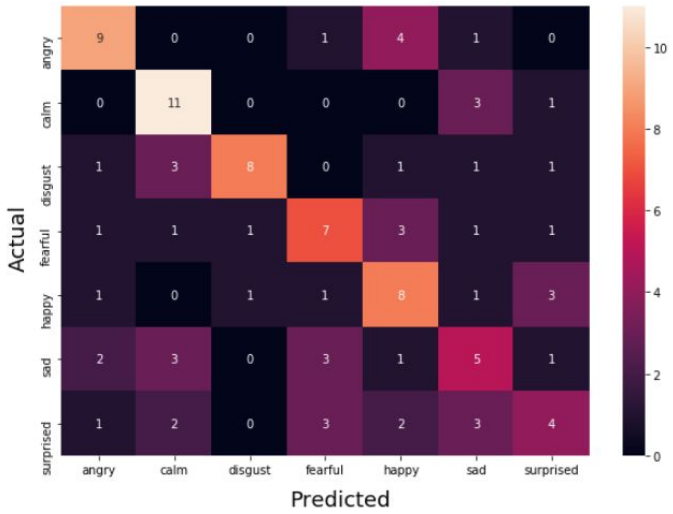
C. Evaluation

After loading back the best model, the training accuracy was 98.934%, the validation accuracy was 51.020% and the test accuracy was 49.524%. The training loss was 0.512, the validation loss was 1.596 and the test loss was 1.773.

The hyperparameters adjusted were L2 regularization rate, dropout rate, optimizer, learning rate, batch size, convolution 1D filter size, max-pooling size, weight initialization.

D. Test

We have created a confusion matrix, to get a better overview from the accuracy in different prediction classes. The best accuracy was achieved on the calm and angry classes, 11/15 and 9/15 accuracy. The worst accuracy was on the surprised and sad classes, 4/15 and 5/15.



VI. Implementation - own CNN version

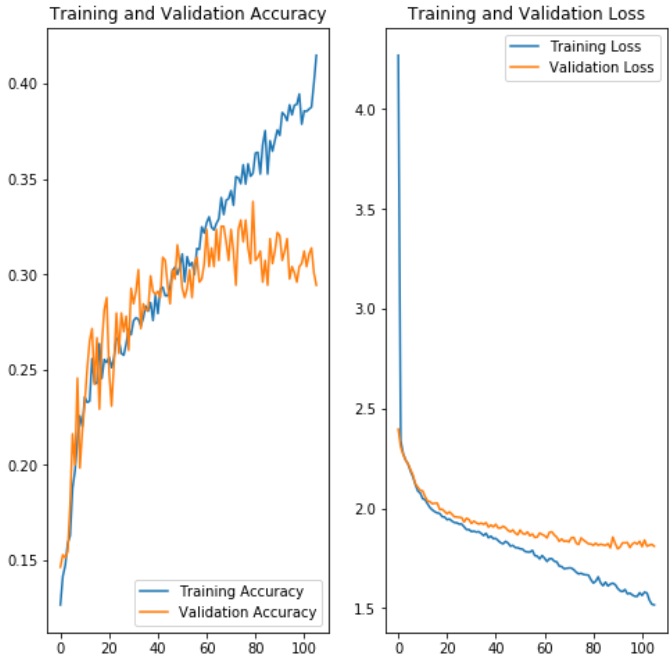
During the development of the model we tried to construct our own neural network for the problem. It was a good experience, and we were able to get a deeper understanding of each process, but it did not achieve really good results.

E. Data preparation

The preparation of the data was similar to the previous procedures. The silent parts of the recordings were cut off, and then spectrogram was made which was proportional to the length of the recording. After that, pieces of the same length became the input images of the model. The input image shape is (128,200,1)

F. Learning - evaluation

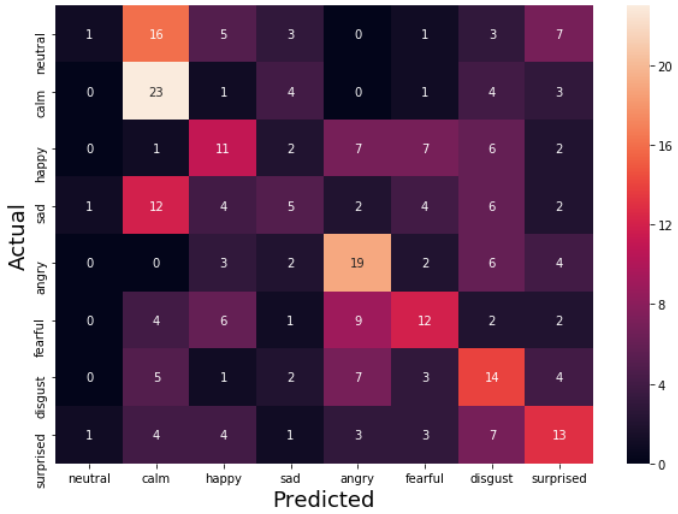
The model is a sequential model consisting of 2-dimensional convolutional layers, maxpooling layers, and dense layers. These have many hyperparameters, and their optimization is a complex process for many layers. However, since I found it best to use few layers, the number of parameters was reduced. Therefore, I did not use hyperparameter optimization methods, but rather parameters were tuned in an intuitive way. The activation function is relu for all internal layers to avoid the vanishing gradient problem. The result is the learning process shown in the following figure.



Additional methods were needed to prevent overfitting.. Dropout sections, L1 L2 regulation were also applied, which allowed the loss value to remain close to each other for train and validation data. However, due to this and the simplicity of the model, the accuracy could not be increased above a value. This value is around 0.35.

G. Test

In the figure below we can better evaluate the results obtained with the help of the confusion matrix. It can be seen that the neutral emotion has not been learned, probably due to the lack of good distinguishing features of this emotion. In addition, our model cannot distinguish sad emotion, but for the other emotions, actual value is the most predictive.



VII. Future plans, summary

To summarize we have investigated three solutions in the speech based emotion recognition topic. We have used the Ravdess data set, which contains sentences said by actors with different emotions. One of the three models used MFCC feature extraction to prepare the data, and the others used spectrogram creation. The best performing model was the MFCC version which achieved 49.5% percent accuracy and 1.773 test loss. Unfortunately, the overfitting prevent us from achieving a better result. We have tried many techniques to prevent it, but the possible problem was that there were not enough training data. As this was our first deep learning problem, we are satisfied

with the results and we have learnt many things during the solution.

References

[1] Reza Chu: Recognizing Human Emotion from Audio Recording. https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3

[2] Dipam Vasani: Sound classification using Images, fastai. https://towardsdatascience.com/sound-classification-using-images-68d4770df426?

[3] Vivek Amilkanthawar: Deep Learning Using Raw Audio Files. https://medium.com/in-pursuit-of-artificial-intelligence/deep-learning-using-raw-audio-files-66d5e7bf4cca

[4] A. Berger, S. Della Pietra, V. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing", Computational Linguistics,

[5] V. Petrushin, "Emotion Recognition in Speech Signal: Experimental Study Development and Application", Proc. Int'l Conf. Spoken Language Processing

[6] B. Schuller, G. Rigoll, M. Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture", Proc. IEEE Int'l Conf. Acoustics Speech and Signal Processing

[7] Julien Epps: Speech Based Emotion Recognition https://www.researchgate.net/publication/283646099_Speech_Based_Emotion_Recognition