Merle Pfau

Speech Emotion Recognition:
LT2316 Machine Learning Course Project

1. Introduction

Spoken language is the most natural form of communication for humans. Therefore, it is no surprise that spoken interaction with machines is long part of our everyday life. But while computers do not struggle with recognizing and generating audio signals, they are as of now unable to understand the underlying emotion in our speech. Speech emotion recognition (SER) is the task of extracting the embedded emotion of a speaker utterance, a task that is often difficult even for humans themselves. Enabling smart mobile devices with this skill would further improve the communication between humans and machines. The key in achieving a well performing system is feature selection, which was also the biggest challenge is the presented project. In this project the goal was to train a deep neural network to classify unseen speech signals according to the emotion conveyed by the speaker. In part 2 of the report I will give an overview on the task of SER, part 3 will explain my data and methods, and in part 4 I will show the results my system achieved. Lastly, in part 4 and 5 I will first discuss the results and used methods and finally give a summary of the project.

2. Speech Emotion Recognition: The task

A central question for SER that is still unanswered is the definition of emotion. One widely used notion is the one of discrete emotions (Ekman 1999). It describes a set of basic emotions, which are culturally independent and natural to humans. These universal emotions are happiness, sadness, surprise, fear, anger, and disgust and are used by people to describe emotions in day-to-day life. In contrast to other definitions of emotion, discrete emotions are experienced for a short period of time. This notion of emotions is widely used in SER systems.

The machine learning task of SER has been around for at least two decades, with solutions based on current technical possibilities of the time, ranging from traditional classifiers to deep neural networks. To this day there is no generally accepted machine learning algorithm for this task. Popular classifiers include Hidden Markov Models, Gaussian Mixture Models, Support Vector Machines and Artificial Neural Networks. In the past years, the performance of deep Neural Networks has improved and surpassed that of the more classical classifiers. A popular choice of NN are LSTMs, because of their ability to remember long term context in the internal memory. This means that LSTMs are suitable for processing sequential data such as time series or speech and a reasonable choice for this task.

Examples of LSTMs being used for SER include Eyben et al. (2010) and Tian et al. (2016) and Kaya et al. (2018), an overview over studies on this task can be found in Akçay et al. (2020).

In preparation of the data for the classification task, several steps of pre-processing are commonly used: Most approaches include some form of Framing, Windowing, Normalization and Noise reduction.

Another main question of the task is the extraction of features. Although many different feature sets have been explored, no consensus has yet been achieved. A common solution is to work with high-dimensional feature vectors based on different categories of speech features. Ayadi et al. (2011) draw a distinction between prosodic or continuous, qualitative, spectral, and Teager-energy-operator (TEO) based features (see Figure 1).

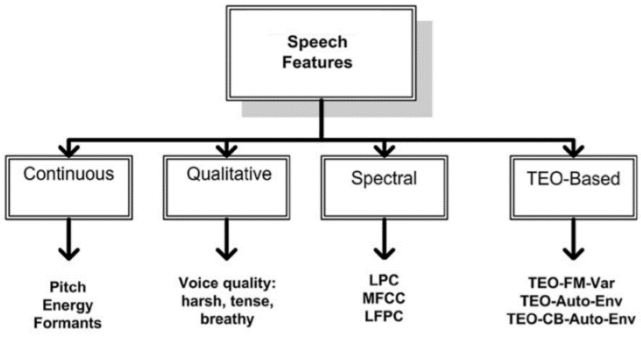

Figure 1: Speech features in Ayadi et al. (2011).

Continuous features such as pitch and energy are widely believed to carry emotional information, which is why they are commonly used in research. Gobl & Chasaide (2003) present evidence on a strong relation between voice quality and the perceived emotion in the hearer. Qualitative features still face some challenges, such as the labels being subjective to the researcher. Spectral features can be described as short-time representations of speech signals: Different emotions have unique distributions of spectral energy across the frequency range of an utterance. There are different algorithms to extract and transform the spectral energy. The last set of features is the TEO-based features based on research by Teager & Teager (1990). They found underlying non-linearities in the vocal tract under stressful conditions. Since hearing is based on detecting the energy of the speech signal, non-linear features are required to accurately represent the speech signal and to detect stress and therefore emotion in the utterance.

On top of this paralinguistic data with audio related features, researchers suggest a multi-modal approach to the task. Several technologies can be used to extract emotions, including visual data, linguistic data (word recognition), or measurements of brain activity.

A central challenge of the task is data collection: There are three different approaches to creating databases of human emotions. The first approach is to collect speech samples of professional actors and actresses that were instructed to perform the emotions. These are typically recorded without any background noise. The problem arising from this approach is the discrepancy to natural data: Firstly,

the emotions acted by professionals do not represent actual authentic human emotions, and secondly in its application the system will most likely be confronted with real life data, containing background noise or even several speakers. A second approach is to try to evoke emotions in speakers and record their utterances to get a more natural portrayal of the emotion. Lastly, there is natural data from e.g. news reports or call centres. There are some legal as well as practical challenges with this approach: Some emotions might not be present in the data, and furthermore the human annotators of the data label 90% of them correctly.

3. Materials and methods

To prepare the data for the task, several steps of pre-processing are taken. The details were inspired by Scott (2020) who explains in detail his approach to solving the task with deep learning, as well as provides the code in a Github repository. Following his approach, I include a pre-emphasis filter described in Damskägg et al. (2019). They suggest that high-frequency content introduced by distortion effects can be difficult to handle for neural networks. Therefore, they introduce a pre-emphasis filter to their model that emphasizes middle and high frequencies in the loss function. It is computed by

$$yt = x_t - \alpha x_{t-1}$$

with $\alpha = 0.97$. Furthermore, before extracting any features, a silence removal step is added. This removes any parts of the speech file that is below 20dB, since these contain little valuable information for the task.

After the data is pre-processed several features are extracted using the library librosa: This step, as well as the data loading follow a tutorial on Speech Emotion Recognition the website data-flair (https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/). The Mel Frequency Cepstral Coefficients (mfcc), a Mel scaled spectrogram (mel), and the pitch (chroma). All these features are computed as global features, more precisely the averages over the whole utterance, which represent the whole sequence rather than sub frames since this seems to be most beneficial for the detection of a majority of the emotions. Studies have proven local features to be beneficial for the classification of the emotions anger and surprise, which are predominantly conveyed at the beginning or end of utterances.

To extract the pitch, the Short-Time Fourier Transform (SFTF) is performed: It calculates the frequency spectrum on sliding frames through the audio file. This SFTF is then used to compute a chromagram using the librosa library implementation feature.chroma_stft. It returns the normalized energy for each chroma bin (pitch class) at each frame, which are averaged over the whole sequence. The feature mel consists of a Mel scaled spectrogram, a magnitude spectrogram that is adapted to the perceived frequency of a hearer. It takes the speech signal and the sample rate, which is the number of samples

of audio carried per second and returns the Mel spectrogram as a numpy array. Again, the average over axis 0 is taken to reduce the number of features. The last feature used in this project are the mfcc: Mel scaled Cepstral Coefficients. Cepstrum is the information of rate of change in a spectrum, so this feature contains information about how the frequency changes within one sample. The computed vector contains 40 coefficients per time frame, which are then averaged per coefficient. This leads to all feature vectors being of the same dimensions and removes the need for additional padding.

These extracted feature vectors are then fed into the LSTM. It consists of two stacked LSTM layers with a dropout of 0.2 after the first layer and a nonlinear ReLU activation layer. The input size in the number of extracted features (180), the output the size of the hidden layer (250). The final state of the last hidden state is then put through a linear layer, which converts it into a tensor of the dimensions (1,8), so one value for each of the target emotion labels. The output of the model is a LogSoftmax probability distribution over the eight labels.

The model is trained for 100 epochs. For training the date is split into batches of size 64. The loss function is the negative log likelihood loss since the output of the model is a softmax distribution. Adam is implemented for the optimizer with a learning rate of 0.001.

The hyperparameters number of LSTM layers, size of the hidden layer, batch size and learning rate were selected through a trial and error process of trying out different combinations. The model with the mentioned hyperparameters achieved the combination of the lowest total loss (3.643) and the highest F1 score (0.6303) when tested with the validation set. The difference to the other models was small though, with most combinations of hyperparameters achieving an F1 score of around 60% in the validation set.

The initial dataset for this project is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). It consists of video, song, and speech samples of 24 actors and actresses perform eight distinct emotions: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. For this project only the 1440 speech samples were utilised: As shown in Figure 1, most emotion are distributed equally, with “angry” being significantly less frequent than the others.

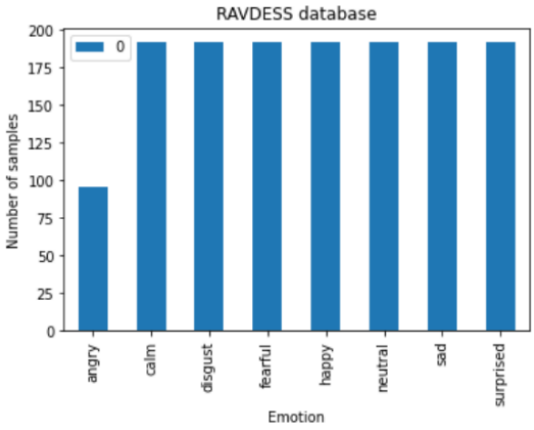

Figure 2: Distribution of emotions in the RAVDESS dataset.

The database was split into 70% training data, 15% validation data and 15% test data, which contain the following distribution of emotions:

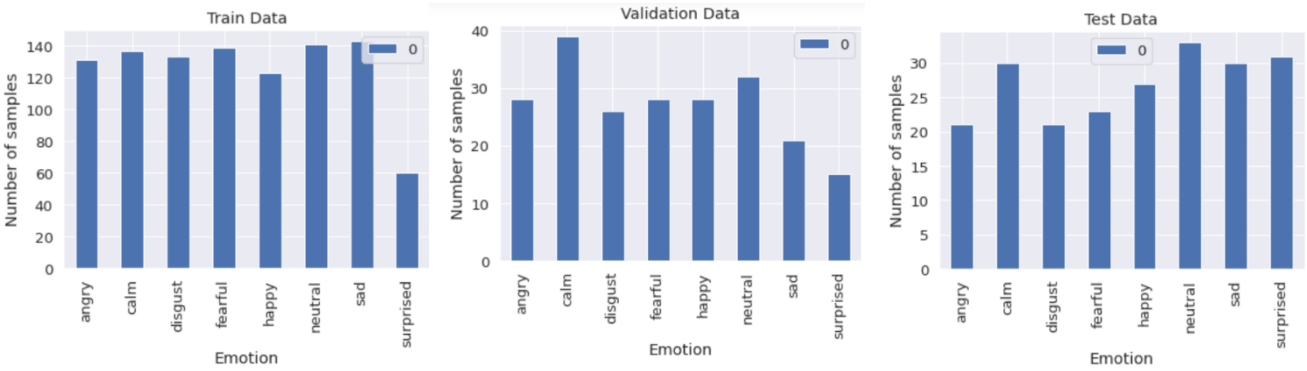

Figure 3: Distribution in the training, testing and validation splits.

To further improve the model's abilities, the same model is afterwards trained on the Berlin Database of Emotional Speech, which consists of 454 professionally acted speech samples of six emotions: neutral, happy, sad, angry, fearful, and disgust. By doing so I hoped that the additional data would improve the abilites of the system, either when classifiying examples from the original RAVDESS dataset, or possibly also result in a working classifier for the Berlin dataset. I was especially interested to find wether the different languages used in the two datasets would influence the outcome, since languages are widely believed to convey emotions differently. Within the Berlin data, it is again split in 70% training, 15% validation and 15% testing data. The distribution of emotions is not quite balanced with some emotions being more frequent in one split than in the others.

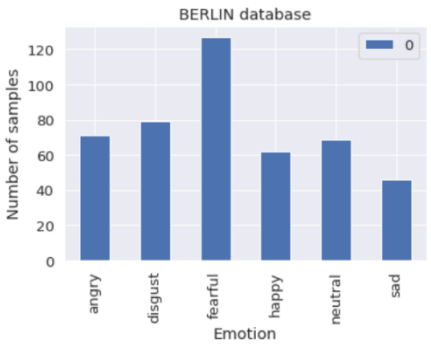

Figure 4: Distribution of emotions in the Berlin Database of Emotional Speech.

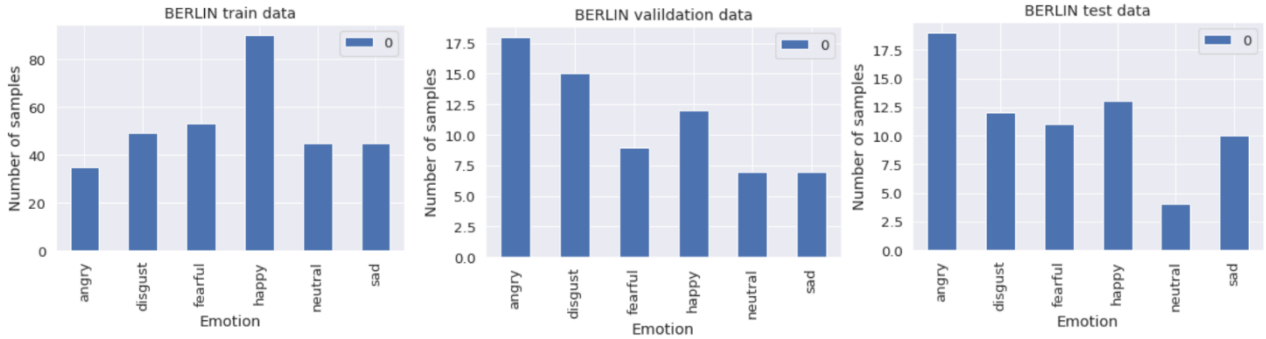

Figure 5: Distribution in the Berlin training, testing and validation splits.

For testing the system, a list of predictions is calculated. Since the output of the model is a probbility distribution over all labels, the prediction is selected by identifying the index of the highest value, corresponding to the id of the most likely emotion. This list of prediciton is then used to calculate the weighted F1 score and to compute a confusion matrix over all emotions.

4. Results

After training the system on the RAVDESS dataset, the system shows reasonable results when tested (see Figure 5). The weighted F1 Score is 0.576. We can see that the model has clearly learned something in the training process, as the results are far from a random distribution. Interestingly, some emotions are classified with a higher precision than others, with fearful and disgust being correctly labelled around 75% of the time and "neutral" and "happy" only 33% of the time.

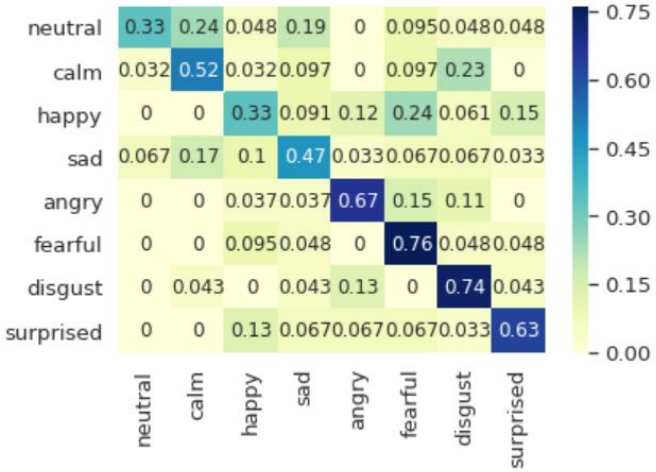

Figure 6: Results after training on RAVDESS.

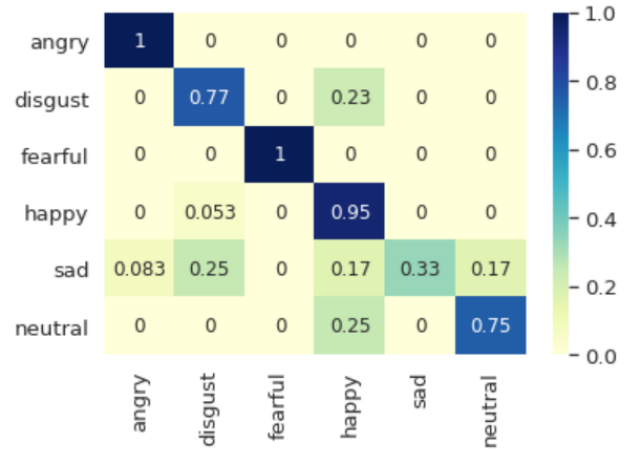

Figure 7: Predictions for the Berlin Database of Emotional Speech.

The system is not able to predict any emotions on a different dataset, the Berlin Database of Emotional Speech before receiving additional training. After completing training on this dataset with the same settings, the results in Figure 6 are achieved, with an overall F1 score of 0.735. As visible in the confusion matrix, the performance on the Berlin dataset is a lot better than previously on the RAVDESS data: For the emotions angry and fearful, all samples were classified correctly. In this experiment, the emotion with the lowest precision is sad, where only a third of the samples received the correct label.

That system with the training on the Berlin dataset performed worse on the RAVDESS data than without the additional training, as can be seen in figure 6.

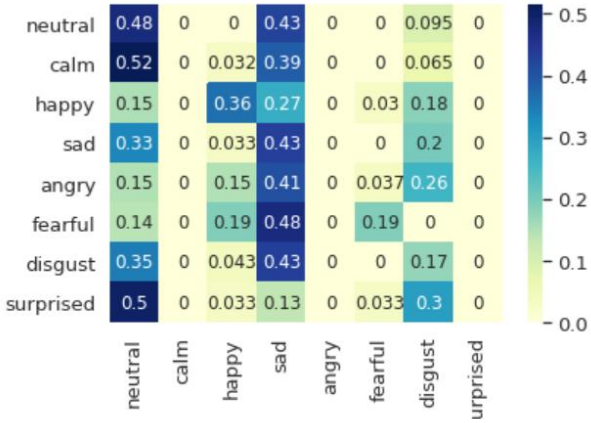

Figure 8: Results on RAVDESS data after training on Berlin dataset.

5. Discussion

Overall, the SER system I created for this course project showed promising results, that could be improved in several different ways. One way would be to include more different databases into the training. Here the selection I made with the combination of RAVDESS and the Berlin database did

not prove to be working, since the present emotions were different in the two databases. One solution is to only combine data that includes the same labels, or more pragmatically only include the emotions of each dataset that are present in all the used data. I would anticipate the results to improve drastically if that step was handled more carefully. Regarding the cross-lingual aspect of RAVDESS consisting English samples and the Berlin dataset being German, I could not see any influence on the results. The decrease in accuracy seems to be solely based on the different sets of emotions, but this would need to be further inspected with a more balanced dataset.

As shown in part 3, the emotions were not distributed evenly in the training and testing data, which might have caused some of the misclassifications.

Furthermore the way of selecting the hyperparameters could have been improved: Adding more different combinations in the selection phase or relying on some external service to provide a set of hyperparameters for the data seems like a valuable addition to the project.

Apart from simply adding more data, some propositions have been made in research that I did so far not include in my implementation but would probably further improve the results.

Luggar & Yang (2008) propose a "2-stage hierarchical classification" model (Figure 8): They separate the (in their case six) emotions into high and low prosodic energy groups. In their classification process they first recognize only the group by looking at prosodic features, and then perform emotion recognition with prosodic & voice quality features within the respective group. This approach could be helpful for my project as well, especially when looking at which emotions were classified with which precision: The lowest values were achieved for the emotions happy and neutral. While utterances with those emotions were mostly classified as members of their group, this was not the case for the third lowest one, calm. In almost a quarter of the utterances the wrong label disgust was assigned, which is high energy opposing the low energy of calm. A pre separations might have helped avoid some of those wrong classifications and could be a future improvement of the system.

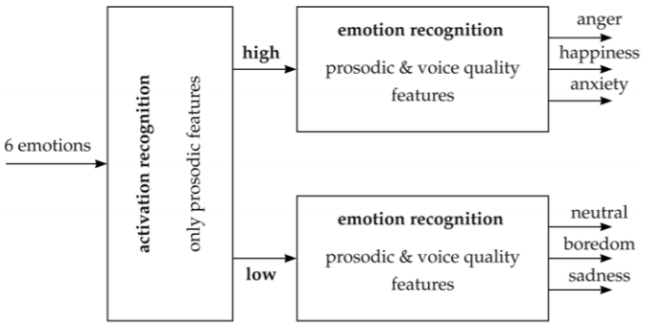

Figure 9: 2-stage hierarchical classification (Lugger&Yang 2008)

6. Conclusion

Overall, the goal of the project was achieved: To train a system to be able to make predictions about the emotional state of the speaker. While the abilities of the system are far from state of the art, it was still able to make decent predictions while only being trained on about 1000-2000 utterances. As

shown in part 4, the system managed to classify unseen speech samples from the Berlin Database of Emotional Speech with an F1 score of 0.735.

Leverage points to improve the results of the project include the selection of features: Including more different features such as voice quality based and non-linear TEO based features would likely further improve the predictions. Furthermore, there were some errors from my side, especially in the corpus selection and combination. I realized too late that the approach of retraining the model on a different dataset with a different set of labels was not going to improve the results. In the future, a combination of datasets with matching categories or rather only training the model on the overlap of different corpora should lead to significant improvement. On top of working on the combination of the two datasets in this project, there is multiple other fitting databases with an overlap in the labels, such as the Crowd-Sourced Emotional Multimodal Actors data-set (CREMA-D) that has a wide variety of age groups and ethnicities in the actors, or the Toronto emotional speech set (TESS). I do still think that the approach of cross-lingual emotion recognition is very interesting and could continue to be of interest in the project.

7. References

Akçay, M. B., Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech communication, January 2020, Vol.116, 56-76.

Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44, 572–587.

Eero-Pekka Damskägg, Lauri Juvela, Vesa Välimäki, et al. (2019). Real-time modeling of audio distortion circuits with deep learning. In Proc. Int. Sound and Music Computing Conf.(SMC-19), Malaga, Spain, pages 332–339.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), Handbook of cognition and emotion. Chichester: Wiley.

Eyben, F., Wöllmer, M., Graves, A. et al. (2010). On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. J Multimodal User Interfaces 3, pp. 7–19.

C. Gobl, A.N. Chasaide. (2003). The role of voice quality in communicating emotion, mood and attitude, Speech Commun. 40 (1–2) 189–212.

Kaya, H., Fedotov, D., Yesilkanat, A., Verkholyak, O., Zhang, Y., & Karpov, A. (2018). LSTM based cross-corpus and cross-task acoustic emotion recognition. In the proceedings of Interspeech, pp. 521–525.

Livingstone SR, Russo FA. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.

M. Lugger, B. Yang, (2008). Psychological motivated multi-stage emotion classification exploiting voice quality features, in: F. Mihelic, J. Zibert (Eds.), Speech Recognition, In-Tech.

Wesley Scott (2020). Deep learning for robust dimensional characterisation of affect in speech. University of Glasgow. https://github.com/terranivium/speech-emotion-recognition

H. Teager, S. Teager. (1990). Evidence for nonlinear production mechanisms in the vocal tract, in: Speech Production and Speech Modelling, Nato Advanced Institute, vol. 55, pp. 241–261

L. Tian, J. Moore and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, 2016, pp. 565-572.