香 港 大 學
THE UNIVERSITY OF HONG KONG

The University of Hong Kong
Department of Computer Science

Final Project

Text Classification and Information Retrieval

Automatic Recognition of Emotions from Recorded Speech Data.

Hanna Dettki
hanna.dettki@student.uni-tuebingen.de

Lecturer: Professor Dr. Yu Lequan

UID: 3035960651, u3596065@connect.hku.hk

Date: April 26, 2022

Contents

Abstract

Natural intercommunication between humans and computers is a key aspect to be achieved in the research of natural language processing (NLP). In order to not only have the machine understand verbal content, but also more subtle signals that any human being would readily react to, inferring emotions from spoken speech in real-time therefore is a fundamental ability for the machine to have.

In this work, two machine learning approaches are evaluated against the task of classifying purely acoustic data as one out of eight possible emotions, namely a Convolutional Neural Network (CNN) and a Support Vector Machine (SVM). It is shown that the SVM significantly outperforms the Deep Learning approach while only taking the most relevant frequency bands into account. Although the latter one is not provided with any temporal context, it predicts emotions on a test set with an accuracy of 63.89% and with an accuracy of 97.05% on the training set.

1 Introduction

Aiming for natural human computer interaction, automatic emotion recognition from speech is a fundamental prerequisite. Automatic speech recognition (ASR) and human auditory comprehension is at the bottom of automatic emotion recognition from speech. It is therefore beneficial to have a good understanding of speech perception and language processing in humans as well as the neurophysiology of the auditory system.

The overall research question underlying various scientists' research in this field is still to gather an understanding of how a speaker's motor gestures are transformed to sounds and how those sounds are mapped onto meaning in the comprehension of spoken language [1]. Despite ASR being a field of research, which is approached by different disciplines of research and there therefore encapsulates a plethora of models such as the classical linguistic one, the cognitive, or engineering one [2], there has been little interaction between those groups according to Moore [1]. While human computer interaction (HCI) control mechanisms previously have relied only on textual or display-based control mechanisms, allowing only the speaker's word-based requests, more intuitive mechanisms are being implemented, exploiting the speaker's voice, gestures, and mimic. The latter ones reveal much more information on the speaker's intentions since more context information is taken into account, such as emotions.

Speech provides a high density of information of which humans percept categories such as *phones* (physical properties of speech sounds) and *phonemes* (systematic organization of sounds). Consequently, it seems to be important to extract such features and map their acoustic properties to emotions in order to provide machines with a better understanding of the human speech. However, in this work, a machine learning approach which only takes the most relevant frequencies into account while losing the temporal context, turns out to work pretty well.

2 Related Work

2.1 Acoustic Properties in Speech

Speech provides a high range and density of acoustical information of which the science of phonetics is concerned with. According to Kuhl et al., the ability to percept phonetic distinctions in a non-native language declines in infants towards the end of the first year. On the other hand, phonetic distinction in the native language improves over time [3]. In a paper by Moore [4], basic aspects of auditory processing involved in the perception of speech is investigated. Frequency selectivity of the auditory system which relates to the ability to resolve sinusoidal components

Table 1: Some variations of acoustic variables observed in relation to emotions, from [8].

Emotion	Pitch	Intensity	Speaking Rate	Voice Qualtiy
Anger	high mean, wide range	increased	increased	breathy; blaring timbre
Joy	increased mean and range	increased	increased	sometimes breathy; moderately blaring timbre
Sadness	normal or lower than normal mean, narrow range	decreased	slow	resonant timbre

of complex sounds, can be evaluated conducting perceptual masking experiments using human listeners. An internal representation of the range of speech sounds in the peripheral auditory system can be derived from the 'auditory filters' inferred from the results of the experiments [4].

Generally, speech recognition depends mainly on the dynamic nature of speech sounds and its change encoded in time [1]. The human auditory system is limited not only by the frequency selectivity but also by the temporal resolution [1]. Therefore, Moore deduces that, for '[...] speech represented in quiet, the resolution of the auditory system in frequency and time usually markedly exceeds the resolution necessary for the identification and discrimination of speech sounds [...]' [1]. Consequently, he concludes that this partly accounts for the robust nature of human speech perception.

Furthermore, Campbell argues that speech perception is multi-modal [5], meaning that what we perceive as speech is not only influenced by what we hear but also by what we see in the facial expression of the talker. This is illustrated by the McGurk effect [6], which is the effect that occurs when an utterance of an audio recording is combined with a video recording of a different utterance. For instance, an acoustic 'mama' is heard as 'nana' if the acoustic 'mama' is paired with a visual 'tata'. This illustrates that what is heard is affected by what is seen which is also shown by Eber [7] who points out that speech is understood in noisy environments much better if the face of the talker is visible to the listener.

2.2 Acoustic Properties of Emotion in Speech

Information on emotion in language is not only encoded in what we say but also in how we say it whereby the 'how' is even more important than the 'what' [8]. Prosody, namely speaking rate, intensity, and pitch are the vocal parameters that are the most important ones related to emotion and which also have been best researched by psychological studies [8]. Table 1 displays seemingly clear acoustic correlates of emotions [8]. These show voice quality and prosody to be the most important voice correlates for human distinction between emotions. Particularly, intensity and pitch appear to be correlated with activation, such that high activation is implied by high intensity and pitch, and low activation is indicated by low intensity and pitch, respectively [8].

2.3 Automatic Speech Recognition

Usually, speech recognition is approached in a two-stage process, in which the acoustic signal is projected onto a series of phonemes in an initial stage, and in a subsequent stage at which the stream of phonemes is segmented into a sequence of words [2]. Correspondingly, much research has focused on linking the acoustic signal's properties to linguistic units such as phonemes and phonological word form representations [2]. However, deep learning has made progress and has therefore enabled end-to-end approaches that directly learn from data and therefore renounce phonemes in their pivotal
role [2].

2.4 Automatic Emotion Recognition

When looking at table 1, automatic recognition of emotions from audio signals seems to be straight forward. However, this is not the case [8]. The psychological studies have to be regarded critically in the sense that their results rely on persons having been told to act in a specific emotional state. In everyday human computer interaction however, such a clear mapping between acoustic variables might be possible in some cases but certainly the variations will be much higher. Wilting and Krahmer show in human listening tests that the perception of acted emotion is different than that from natural emotions [9]. In addition, acted emotions are judged stronger than natural emotions by listeners in a perception experiment [10]. Obviously, natural emotions are of greater interest for HCI. Adversely, automatic emotion recognition becomes more complex the more natural the database is, which is accounted for by the mapping of acoustic variables which become less correlated the more natural emotions are [8]. According to Vogt et al., the labeled emotions in a dataset and consequently the ones that may be recognized, can consist of a set of the classic emotions such as sadness, disgust, anger, and joy or, alternatively, emotions can be placed into a dimensional model comprised of two or three affective states. Usually, the dimensions are arousal (from high to low) and valence (from positive to negative) [8]. For automatic emotion recognition, the dimensional model is usually mapped onto the four quadrants (see figure 1). Also, there are only a few datasets available that are labeled using a dimensional model [8] which makes it particularly hard to train a classifier based on this model, especially if the dataset is not freely available.

2.4.1 Applications

Generally, there is a wide range of applications in which emotion-aware speech recognition systems may be of use. For instance, in a call center, such an application could sort the incoming voice messages according to the emotion conveyed by the caller and could hence select the appropriate conciliation strategy. Furthermore, the automobile industry is researching how the emotional state is influencing the driving behavior [8] and may develop corresponding driving assistants. Lastly, real-time emotion recognition in HCI-applications, for instance in the context of a nursing robot, needs yet to be mastered. In this context, the term 'emotion' may thus be interpreted in various ways and would hence comprise all sorts of human affects that may occur in HCI. Nevertheless, the more realistic the data becomes, the less feasible it is to classify emotions and human affect states correctly [8]. Figure 2 shows how the difficulty of emotion recognition correlates with the dataset being used.

2.4.2 Automatic Speech Emotion Recognition System

As depicted in figure 3, a speech emotion recognition system consists of three principal parts: signal processing, feature extraction, and classification. Signal processing involves acoustic pre-

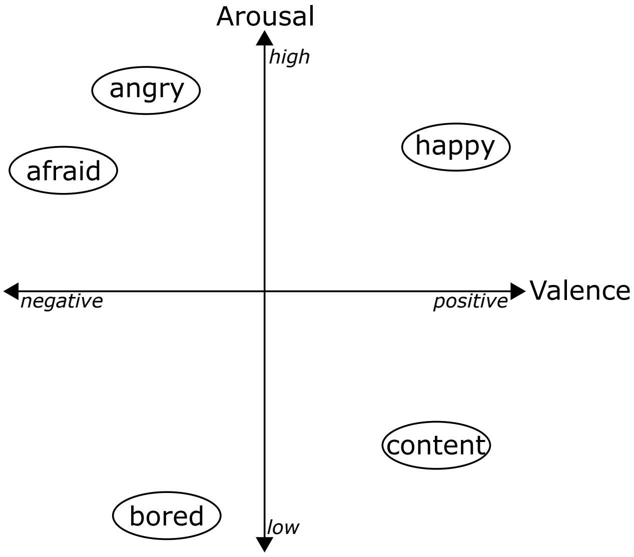

Figure 1: A two-dimensional emotion space with an Arousal and a Valence axis. Basic emotions are marked as ellipses within the quadrant. Adapted from [8].

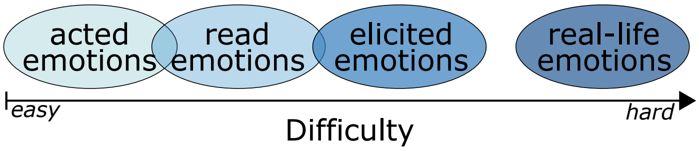

Figure 2: Types of databases used for emotion recognition and their difficulty. Adapted from [8].

processing, such as filtering the samples, as well as potentially segmenting them into meaningful units. Feature extraction is concerned with identifying the relevant features of the acoustic signal that are representative for emotions. Lastly, classification maps feature vectors onto emotion classes through learning by examples [8].

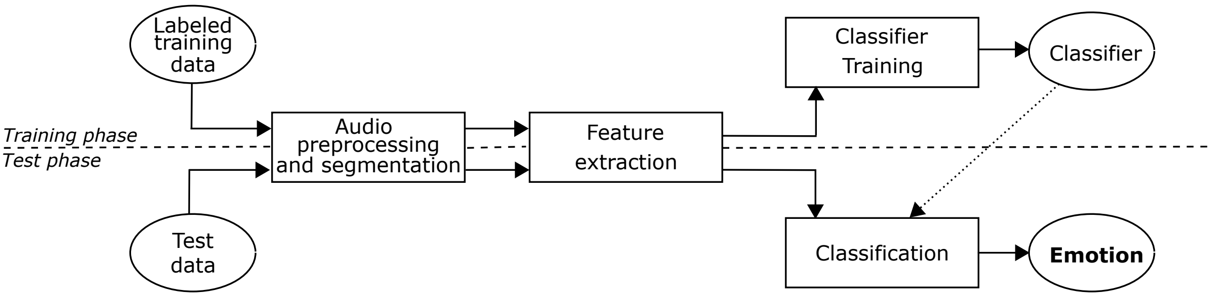

Figure 3: Overview of speech emotion recognition system. Adapted from [8].

3 Data Processing

3.1 Signal Processing

3.1.1 Audio Segmentation

Sound is transmitted in one-dimensional sound waves which have a single value based on the wave's height at every moment in time. Next, sampling needs to be done which is the process of turning the sound wave into bits by recording the height of the wave at equally-spaced points. While for 'CD Quality' audio is sampled at 44.1 kHz, a sampling of 16 kHz is enough for speech recognition in order to cover the frequency range of human speech. According to the *Nyquist criterion*, the original sound wave can be reconstructed from the sampled one, as long as it is sampled at least twice as fast as the highest frequency aimed for to be recorded [11]. For this reason, no information is lost when sampling.

3.2 The RAVDESS Dataset

In this work, the *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) is used. It is free of charge and can be downloaded at `https://zenodo.org/record/1188976`. The RAVDESS database 'is a validated multimodal database of speech and song [...]' [12]. 24 professional actors and actresses, equally balanced in gender, were recorded phrasing main clauses such as 'Dogs are sitting by the door.' in North American English in one of the following eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised. Each of the 7356 RAVDESS files corresponds to a unique file name of which only the 1440 (= 60 trials per actor x 24) audio-only speech files are used in this work.

3.3 Investigating the Data

First, the raw audio waves as .wav data are loaded into the python script. As mentioned in section 3.3.1, this data could now be fed into the neural network, but speech or emotion recognition turns out to be difficult on raw audio waves. Thus, some more preprocessing is done. Thereby, two problems have arisen, namely the variable length of the raw audio data, since CNNs require input samples of equal dimensions and the fact that the recordings consist of unnecessary and meaningless silence before and after the actual recorded speech as depicted in figure 4. Hence, the audio samples are cropped in order to get rid of the meaningless silence. However, the trimmed audio files still are of different lengths, as shown in figure 5. This particular problem is solved differently for the two approaches and is thus elaborated on in more detail in the respective sections.

The raw audio wave data is transformed into a spectrogram as described in section 3.3.1. Figure7 depicts the corresponding spectrogram to the raw audio wave shown in figure 4 and figure 7 depicts the respective spectrogram on the trimmed audio wave.

3.3.1 Audio Preprocessing

Arrays of numbers represent the amplitude of the sound wave at $\frac{1}{16000}$th of a second interval. This data could now be fed into a neural network, but recognizing speech or emotion patterns by processing these samples turns out to be difficult [13]. Instead, some more preprocessing on the audio data can be done by breaking apart the complex sound into its component parts by applying the Fourier Transform which breaks apart the complex sound wave into the simple sound waves that it is made up of [14]. Essentially, the sound wave is now split into its frequency parts. By adding up the energy that is contained in each frequency band, a quantification of the relevance of each frequency is obtained at every point in time. This representation is called a *spectrogram* which is a suitable input representation of audio data for a neural network whereby the intensity of the spectral coefficients is plotted versus the time index.

4 Methods

In this project, two approaches are investigated: first, a *Convolutional Neural Network* (CNN) (see section 4.1) and second, a *Support Vector Machine* (SVM). Having computed the spectrograms from the audio data, we now have appropriate features available that can be processed by a learning algorithm. As the first approach, an CNN is trained on a spectrogram (see figure 7). However, neural networks require a large number of data which should exceed the 1440 samples of our data set by orders of magnitude. Thus, a second approach using a support vector machine is investigated since SVMs can handle small data sets well.

For both approaches, 10% of the data set is used as independent test set to evaluate the model's ability to generalize.

Below, a very brief overview of CNNs and SVMs is given from a high-level perspective.

4.1 Approach 1: Convolutional Neural Networks

4.1.1 Foundations CNN

Whilst being inspired by the concept of the mammalian retina, Convolutional Neural Networks (CNN) have been proven to be outstandingly effective on high-dimensional input data [14]. CNNs are a special kind of neural network, which use convolution as a specialized kind of linear operation in place of a general matrix multiplication [14]. Advantageous to CNNs is their high efficiency in terms of computational complexity while using a sparse set of parameters [14]. The kernels which the CNN learns during training process, are reused over the entire input which markedly exceeds regular feed forward networks in terms of memory and computational efficiency. For further details on the topic, the reader is referred to [14].

4.1.2 Implementation of CNN

The audio data is processed with a Fourier Transform whereas the audio waveform can be transformed into a tensor of rank two, i.e. a matrix. In doing so, the rows of the matrix correspond to different frequencies and the columns relate to different points in time. However, the spectrograms are still of different lengths which is why zero padding is used to transform all samples to equal lengths. Thereafter, a min-max scaling is applied which means that to every value, $x = \frac{(x-min)}{max-min}$ is applied which normalizes the data to a value range of $[0, 1]$. Hereby, min and max are the respective minimum and maximum values over the entire dataset.

4.1.3 Evaluation of CNN:

Surprisingly, in contrast to the prior assumption, the chosen model seems to be too simple and does not manage to learn the problem from the data. In every experiment ran – due to the restricted computational power of the notebook used – the training error does not converge, but rather oscillates around a constant value. It is safe to assume that the considered models are too small and thus not powerful enough to make sense of the data. Sine the training error is already unsatisfying, the model is not evaluated on the test set.

4.2 Approach 2: Support Vector Machines

Support Vector Machines (SVM) are a widely used classifier due to its flexibility in modeling diverse sources of data and its ability to deal with high-dimensional data [15]. The goal of SVM is to find a hyperplane of the form $\mathcal{H} = \left\{x \in \mathbb{R}^d | \langle w, x \rangle + b = 0\right\}$ such that it separates the data while maximizing the distance between the hyperplane and the closest data point. This distance is referred to as the margin (see figute 9). By maximizing the margin, the classifier is most robust to noise in new data [16].

4.2.1 Implementation of SVM:

For the SVM, the Python module *Scikit-learn* is used [17]. In order to feed the SVM a proper feature vector, the spectral coefficients from the spectrogram are summed up over time. While integrating over the time dimension, the temporal context is lost completely, however.

Note that for this approach, no temporal dependencies in the audio data are taken into account, which seems counter-intuitive for speech processing at first glance. The model is expected to infer the emotions solely based on the frequency distribution as depicted in figure 8. The figure shows both – before and after centering – the cumulative sum (y-axis) over each considered frequency band (x-axis) from the spectrogram. This results in a 'frequency-histogram' on which a dimension reduction algorithm, namely *Principle Component Analysis* (PCA), is applied to reduce the dimensions of the input data in order to obtain fewer, more meaningful features [16]. Thereby, the optimal parameters are obtained via Grid Search, both for PCA and SVM.

5 Conclusion

Two approaches have been investigated for the task of automatic emotion recognition. The motivation was to use machine learning algorithms in order to infer emotions from recorded speech. Thereby, a Deep Learning approach was chosen first, whereby a Convolutional Neural Network (CNN) was trained on the RAVDESS data set. Regrettably, it did not learn the problem at all. Due to the small number of training samples and the lack of computational power, this approach was not further pursued but certainly remains an interesting topic to investigate within a larger scaled project. Therefore, the insights of the first attempt gave rise to a different approach, namely a Support Vector Machine (SVM). An SVM works well on smaller data sets and delivered surprisingly good results. An accuracy of 97.05% on the training set and 63.89% on the test set was achieved. This demonstrates that the SVM predicts emotions correctly way beyond chance and in addition, only based on most relevant frequencies related to the respective emotion since the temporal context is lost completely in this latter approach. This suggests that this machine learning algorithm is able to predict emotions from recorded speech data only based on the respective, relevant frequencies. Consequently, one may conclude that emotions are encoded in specific frequencies to a large extent. Undoubtedly, these results are far from being able to be generally applicable for automatic emotion recognition, since this task becomes more difficult the more realistic the data set is, which is illustrated in figure 2. Unfortunately, it is difficult to compare results on automatic emotion recognition among publications since most researchers use different speech data sets, speaker types, and classifiers. Additionally, in this field of research, there exists no consistent benchmark, such as MNIST in the field of pattern recognition on images [18].

5.1 Future Work

Besides these already promising results, further investigation regarding the deep learning approach might be worthwhile. For instance, different neural networks such as Recurrent Neural Networks (RNN) [19] or a combination of different architectures might be interesting to investigate. Additionally, a different padding to the spectrogram could be applied. Furthermore, different parameters could be tried with the SVM. Moreover, a dynamic classifier such as Hidden Markov Models (HMM) seems to work well. Such dynamic modelling techniques are also the most common ones found in literature on automatic emotion recognition on speech [8].

Overall, this project shows, that SVMs are a promising technique for the task of automatic emotion recognition on speech data.

6 Appendix

6.1 Figures of Data Processing

6.1.1 For CNN

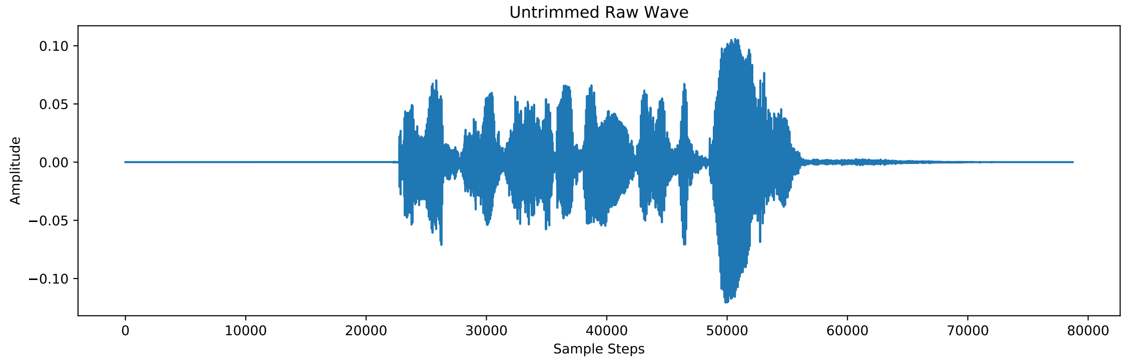

Figure 4: Un-trimmed sound wave of woman saying 'Dogs are sitting by the door.' in an fearful manner. There is still a lot of meaningless silence before and after the woman speaking.

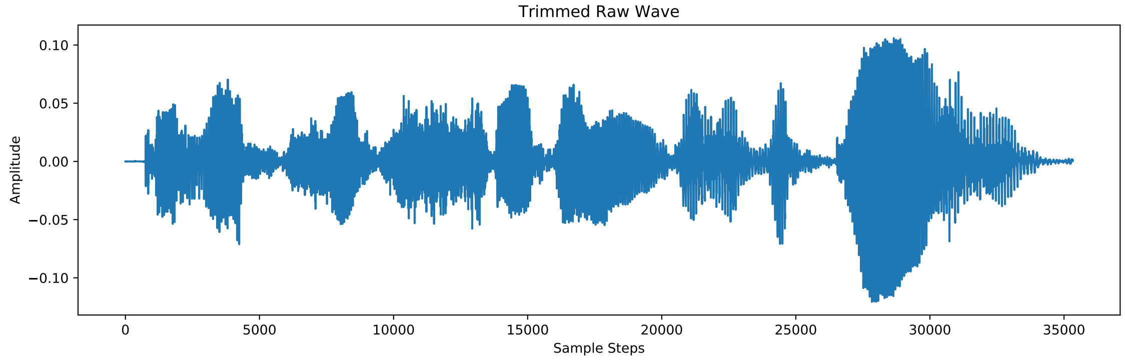

Figure 5: Corresponding trimmed raw wave which now only contains the recorded speech since the silence was trimmed.

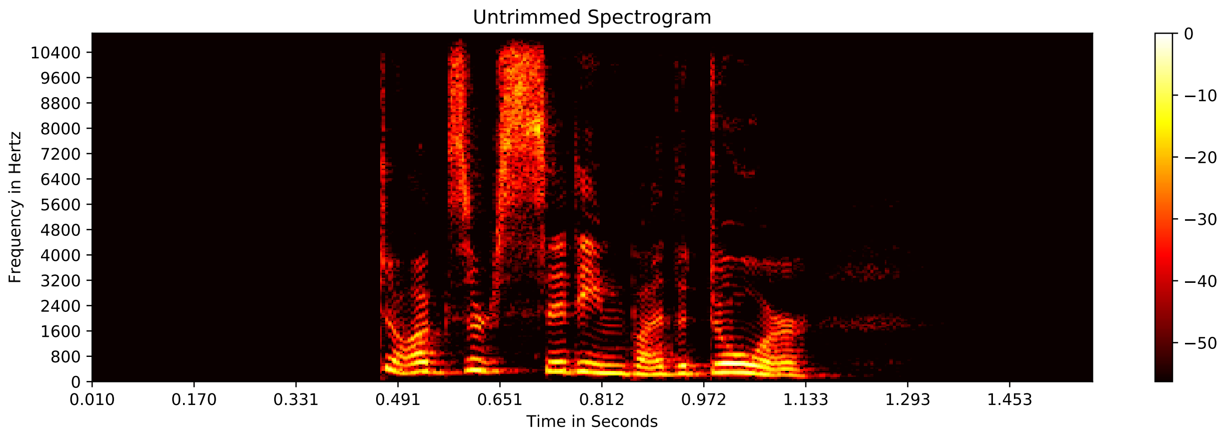

Figure 6: Corresponding spectrogram of un-trimmed raw wave of woman saying 'Dogs are sitting by the door.' in a angry manner.

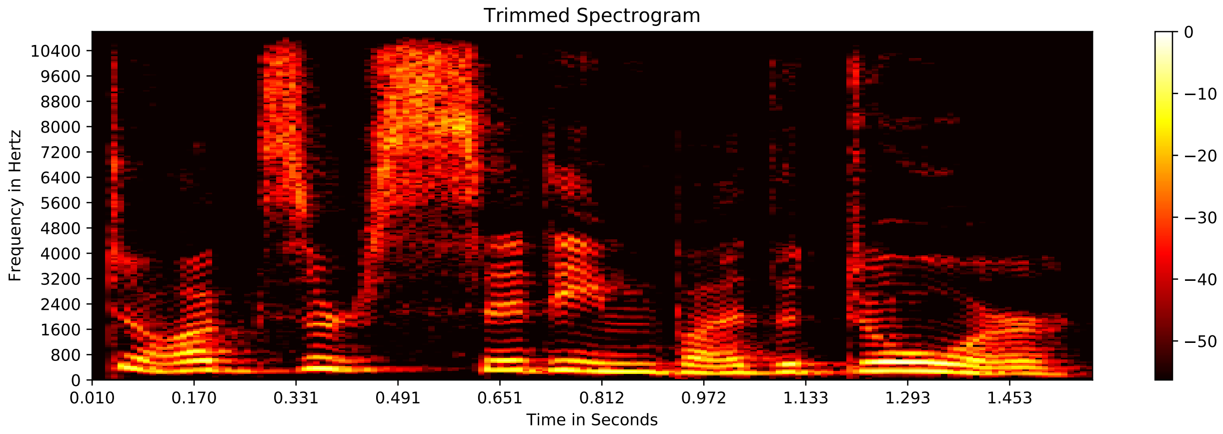

Figure 7: Respective spectrogram of trimmed raw wave.

6.2 Figures and Parameters for SVM-Model

6.2.1 SVM Model Parameters

From the Grid Search, the following values are obtained as optimal parameters:

$$\begin{aligned} l &= 745 && \text{(number of principal components of PCA)} \\ \gamma_{\text{PCA}} &= 1 && \text{(inversed kernel width of PCA)} \\ C &= 10 && \text{(soft-margin constant of SVM)} \\ \gamma_{\text{SVM}} &= 10 && \text{(inversed kernel width of SVM)} \end{aligned}$$

Note hereby, that the PCA is applied in an n-dimensional feature space where n is the number of samples, and therefore resulted in a dimension higher than the data dimension. Further details on this topic are beyond the scope of this work but can be looked up in [20]. Furthermore, it is noteworthy, that there might exist better hyperparameters than the ones Grid Search has determined in our case.

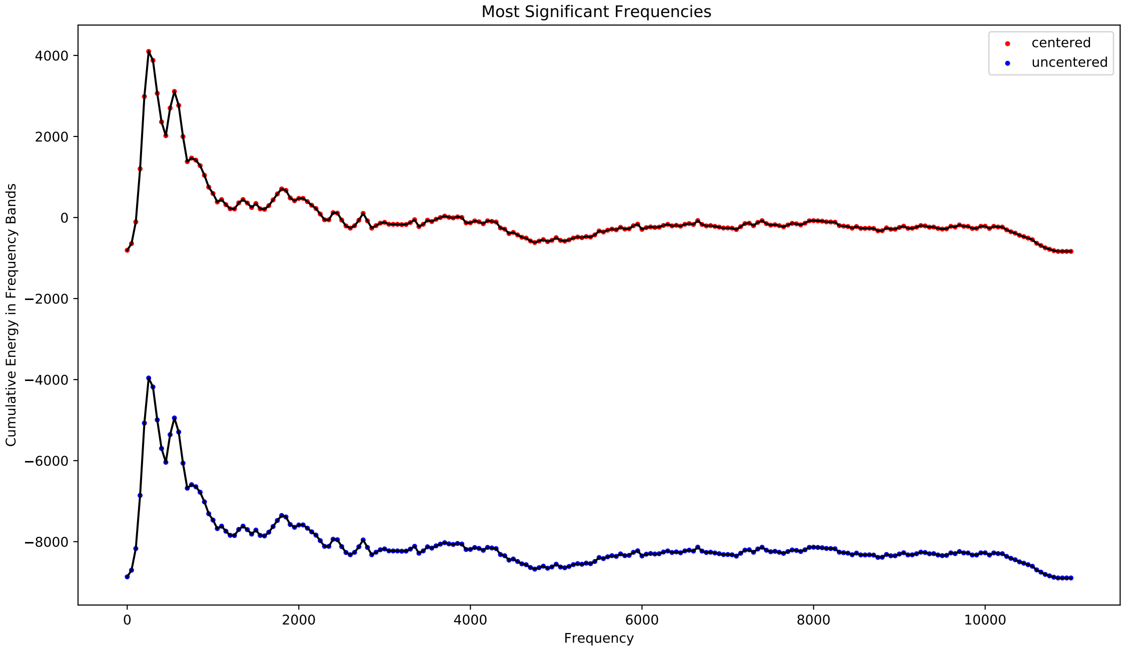

Figure 8: The cumulative sum of the frequencies in the respective frequency bands as input for the SVM.

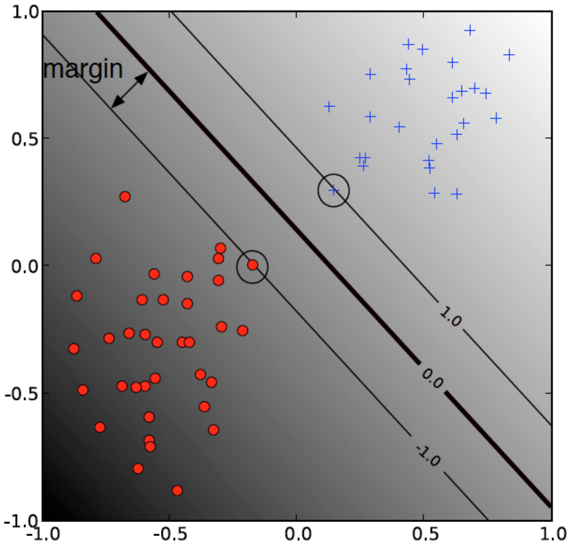

Figure 9: A linear SVM. The circled data points are the *support vectors* – i.e., the examples that are closest to the decision boundary. The support vectors determine the the margin with which the two classes are separated. In this work, eight classes need to be separated. From [15].

7 References

References

[1] Brian CJ Moore, Lorraine K Tyler, and William Marslen-Wilson. Introduction. the perception of speech: from sound to meaning, 2008.

[2] Elnaz Shafaei-Bajestan and R Harald Baayen. Wide learning for auditory comprehension. *Proc. Interspeech 2018*, pages 966–970, 2018.

[3] Patricia K Kuhl, Barbara T Conboy, Sharon Coffey-Corina, Denise Padden, Maritza Rivera-Gaxiola, and Tobey Nelson. Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493):979–1000, 2008.

[4] Brian CJ Moore. Basic auditory processes involved in the analysis of speech sounds. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493):947–963, 2008.

[5] Ruth Campbell. The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493):1001–1010, 2008.

[6] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976.

[7] N. P. Eber and H. Birk Nielsen. Hearing lips and seeing voices. *In Visual and audio-visual Perception of Speech*, pages 12–30, 1976.

[8] Thurid Vogt, Elisabeth André, and Johannes Wagner. Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In *Affect and emotion in human-computer interaction*, pages 75–91. Springer, 2008.

[9] Janneke Wilting, Emiel Krahmer, and Marc Swerts. Real vs. acted emotional speech. In *INTERSPEECH*, 2006.

[10] Emmett Velten Jr. A laboratory task for induction of mood states. *Behaviour research and therapy*, 6(4):473–482, 1968.

[11] Walt Kester. What the nyquist criterion means to your sampled data system design. *Analog Devices*, pages 4–6, 2009.

[12] Steven R Livingstone and Frank A Russo. 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, 2018.

[13] Adam Geitgey. How to do speech recognition with deep learning. 2016. URL `https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c`

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[15] Asa Ben-Hur and Jason Weston. A user,Äôs guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer, 2010.

[16] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006. URL `http://scholar.google.com/scholar.bib?q=info:jYxggZ6Ag1YJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0`.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] Yann LeCun. Mnist database of handwritten digits. URL http://yann.lecun.com/exdb/mnist/.

[19] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385 of *Studies in Computational Intelligence*. Springer, 2012. ISBN 978-3-642-24796-5.

[20] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.