Recognition of Emotional Speech with Convolutional Neural Networks by Means of Spectral Estimates

Conference Paper · October 2017

DOI: 10.1109/ACIIW.2017.8272585

Ronald Böck
Otto-von-Guericke-Universität Magdeburg

Andreas Wendemuth
Otto-von-Guericke-Universität Magdeburg

Recognition of Emotional Speech with Convolutional Neural Networks by Means of Spectral Estimates

Norman Weißkirchen
Cognitive Systems Group
Otto von Guericke University Magdeburg
Magdeburg, Germany
Email: norman.weisskirchen@ovgu.de

Ronald Böck and Andreas Wendemuth
Cognitive Systems Group, Otto von Guericke University &
Center of Behavioral Brain Sciences Magdeburg
Magdeburg, Germany
Email: {ronald.boeck, andreas.wendemuth}@ovgu.de

Abstract—**Current developments in deep neural architectures achieved remarkable results in the classification of emotions from speech. Recently, also cross-modal approaches gained attention in the community. Such a classification method is the Convolutional Neural Network (CNN). Mainly developed for analyses of images it can be used also in speech processing. In this paper, we present a CNN-based classification architecture adapting spectrograms as representations of emotion-afflicted speech input. Given this approach, we applied our network architecture to three benchmark corpora, namely EmoDB, eNTERFACE, and SUSAS, and investigated the classification ability in a Leave-One-Speaker-Out setting. Especially, for SUSAS, a close-to-real-life corpus, remarkable results were obtained. In addition, we investigated the option of analysing CNN's internal representations of the given input using Deep Dreaming. For this, we were able to identify spectral parts which contribute most to the classification process.**

1. Introduction

Considering the current progress in classifiers used for evaluations in various modalities we can state that recently deep neural networks (cf. for speech modality e.g. [1], [2]) gained lots of attention. In addition, such networks are also used in the emotion recognition from speech (cf. e.g. [3], [4]). Especially for the interpretation of visual affects a deep network architecture shows remarkable results (cf. e.g. [5]), namely the Convolutional Neural Network (CNN).

In the current paper, we investigated CNNs for the classification of emotions from speech. To the best of our knowledge, currently just a few groups are working on this issue using similar inputs, namely spectrograms as representation of speech samples, but report on different architectures. More details will be discussed in Section 2. In particular, in CNNs a spectrogram represents the visual equivalent of acoustic samples preserving the emotional content given in the audio material (cf. [6]). For this, CNNs can be used on acoustic samples.

As we will see in Section 2 most of the work on emotion recognition from speech applying CNNs is done on EmoDB (cf. [7]). In our study, we extended these investigations to two other benchmark corpora (cf. [8]), namely eNTERFACE (cf. [9]) and SUSAS (cf. [10]) showing that the proposed architecture (cf. Section 4.2) can be used corpus-independently providing material with different characteristics. Mainly the material is recorded in different environments, reflecting samples with acted or spontaneous speech and emotions. Therefore, covering a variety of acoustic characteristics we demonstrated the suitability of CNN-based recognisers in emotion recognition from speech. The analyses presented in this paper are based on the work of [11] where the architecture is developed and tested on EmoDB. Given the neural network, as discussed in Section 5.1, we achieved similar results as the other working groups on EmoDB and for the two additional datasets we obtained comparable results (average unweighted recall for SUSAS 0.57 and for eNTERFACE 0.66) related to published achievements with different classifier approaches (cf. [8]).

Additionally, in CNNs a kind of an internal representation of the learned content can be visualised showing how each network interprets the input. This approach by Google is called *deep dreaming* (cf. [12], [13]) and is especially used to highlight parts of input images for which the CNN is sensitive. For instance, in a picture showing a meadow a group of trees could be emphasised since the particular network is sensitive towards any structure that can be interpreted as trees. In addition, this method can also be used to investigate and interpret the internal representation of patterns in CNNs during a classification process. Therefore, we analysed the trained CNNs on EmoDB with deep dreaming to generate or recreate the internal representation of the given spectrograms showing 1) that the input is mapped appropriately and 2) that apparently the networks focus on some frequencies in the emotion classification (cf. Section 5.2).

2. Related Work

CNNs – being a specific architecture of deep neural networks (cf. e.g. [2]) – are a current state-of-the-art method which combine several layers containing different functional units to achieve a higher complexity in pattern recognition tasks. Originally developed for image recognition (cf. e.g.

978-1-5386-0680-3/17/$31.00 ©2017 IEEE

[13]) we will adapt the approach in our experiments for processing acoustic data (cf. also [14]).

In the CNN-related community several architectures have been investigated. Currently, mainly five basic approaches are used, namely the LeNet (cf. [15]), the AlexNet (cf. [16]), the ZFNet (cf. [17]), the GoogleNet (cf. [13]), and the ResNet (cf. [18]). In our experiments we focussed on the architecture related to the AlexNet (cf. Section 4.2) since it won the 2012 ImageNet Large Scale Visual Recognition Challenge achieving an error rate of 15.4% (cf. [5]).

In Section 4.1 we discuss the data preparation process of acoustic material to be used in CNNs. For this, we applied spectrograms as a visual representation of an audio input. As shown in the literature (cf. e.g. [6], [19], [20]), spectrograms allow a transformation of acoustic input into a visual representation without loss of information. This approach was already tested in speech recognition tasks (cf. [14], [21]). In the context of emotion recognition from speech it is important to preserve the emotional characteristics (e.g. prosodic information in emotion recognition; [22]).

In particular for the current paper, four publications are important to be considered in more detail:
The mentioned works have in common that they use spectrograms as representations of audio inputs and that the presented methods were tested on, at least, the EmoDB corpus (cf. [7]). As this dataset is a prominent benchmark corpus in the field of emotion recognition from speech (cf. e.g. [8], [23]) we used amongst others EmoDB in our experiments as well.
In [20], the authors present a framework for emotion recognition from speech using a combination of CNNs and Support Vector Machines (SVMs). The speech input is transformed into a spectrogram and presented to the CNN. In contrast to the work discussed in the current paper, the CNNs just extract features from the spectrogram which are used in the classification based on SVMs. Further, the classification was tested on three other datasets as well. To the best of our knowledge, the authors do not use a sequence of spectrograms.
The following three described works have been done in parallel in different groups considering CNNs as classifiers for emotion recognition from speech. The authors of each publication did not know about the others works, therefore, the main ideas – using spectrograms in combination with CNNs – are similar but the particular realisations and results are quite different. Recently, in [24] a combination of CNN layers with recurrent layers are tested on EmoDB (cf. [7]). Also recently, in [25] a combination of three CNN layers and three full-connected feed-forward networks (cf. Figure 2 in [25]) are used to predict emotions given in the EmoDB corpus (cf. [7]). Different sizes of convolutional layers have been applied, namely 120, 256, and 384 units, respectively. The authors extracted roughly 3000 spectrograms from the whole corpus reporting that they thus have 500 images per emotion. For the evaluation, a cross-validation paradigm is selected. The setup is tested only on EmoDB (cf. [7]).
In [11] a similar approach as in [25] is investigated. Also, a combination of CNN and feed-forward layers is applied to EmoDB (cf. [7]). In contrast, the number of layers varies as well as the order of certain layers. Both works were done in different groups at the same time considering different goals. The main aspect of [25] is on prediction whereas in [11] the focus is on a classification of emotions in relation to deep dreaming. Further, both works differ in the way of handling the original speech input. This work ([11]) provided the foundation of the current paper's research.

3. Corpora

In the experiments related to emotion recognition from speech using CNNs we utilised three corpora (cf. Table 1), known in the community, serving as benchmark data sets (cf. e.g. [23]). We decided for a variation of quality conditions showing the flexibility of our approach (cf. Section 4.1).

3.1. EmoDB

The *Berlin Emotional Speech Database (EmoDB)* (cf. [7]) is a well-known studio recorded corpus providing audio samples from ten (five female) professional actors. They utter ten German sentences with emotionally neutral content in a noise- and echo-reducing audio cabin. EmoDB contains 535 phrases in seven predefined categories of emotional expressions: anger, boredom, disgust, fear, joy, neutral, and sadness. The corpus' material was selected using a perception test (cf. [7]). In general, EmoDB represents a set of copora containing audio samples with expressive emotions in high acoustic quality.

3.2. eNTERFACE

The *eNTERFACE* (cf. [9]) corpus comprises recordings from 43 subjects (eight female) from 14 nations. It consists of office environment recordings of pre-defined spoken content in English including accents and dialects. Overall, the data set consists of 1277 emotional audio instances in six text-induced emotions: anger, disgust, fear, joy, sadness, and surprise. Although the affective material is acted, the quality of emotional content spans a much broader variety than in EmoDB, especially in term of echo and noise. Thus, eNTERFACE represents a corpus containing acted material with non-ideal recording conditions.

3.3. SUSAS

Speech Under Simulated and Actual Stress (SUSAS) (cf. [10]) is a data set containing both spontaneous and acted emotional instances where the speech signal is partly masked by field noise. Therefore, especially in terms of spontaneous speech, we assumed to apply acoustic data in close-to-real-life conditions. In particular, we chose a corpus' subset providing 3593 actual stress speech segments recorded in fearful and stressful tasks (cf. [8]). Seven subjects (three female) in roller coaster and free fall stress situations utter emotionally coloured speech in four categories: neutral, medium stress, high stress, and screaming.

TABLE 1. OVERVIEW OF THE SELECTED EMOTION CORPORA USED IN THE EXPERIMENTS.

Corpus	Content	All	Subjects	Emotion	Quality	Audio channel	Length (HH:MM)
emoDB	German fixed	535	10 (5 female)	acted	studio	16 kHz 16 bit	00:22
eNTERFACE	English fixed	1277	42 (8 female)	acted	noisy	16 kHz 16 bit	01:00
SUSAS	English fixed	3593	7 (3 female)	mixed	noisy	8 kHz 16 bit	01:01

4. Experimental Setup

The experimental setup section provides information on the preprocessing of the given audio data and further on the system's architecture. Additionally, we briefly present the validation paradigm and measures used in the experiments.

4.1. Data Preparation

The most important step in the data preparation is the preprocessing of audio data to fit the prerequisites of CNNs. As already discussed, such networks are powerful in processing visual inputs. Therefore, the idea was to transfer the acoustic information into a visual representation. For this issue, we applied the spectrogram which contains the information from the original recording, in particular frequency and amplitude, preserving the emotional characteristics coded in prosodic features (cf. e.g. [8]).

To obtain the spectrogram of a given audio snippet[1] we processed the data as follows: We chose a Hamming window with a window size of 20 ms as commonly used in the speech processing community. On the one hand, this ensured that at least one phoneme – carrying also emotional information (cf. [22]) – is included in the particular window. On the other hand, the window is large enough to cover at least two periods of the lowest basic frequency within human speech (around 100 Hz) [26]. We then applied a Fast Fourier transformation on the signal. An example of the final spectrogram is presented in Figure 1. The transformation to a spectrogram allows to focus on the affective content of the uttered speech compared to the waveform which additionally varies regarding the spoken content. This is assumed to reduce the confusion for a CNN-based classifier.

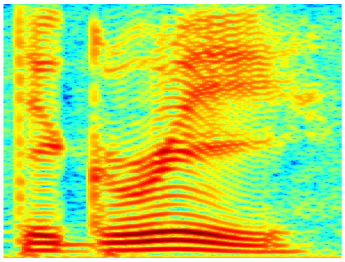

Figure 1. Example from the SUSAS Dataset (Speaker: M2, word: degree)

As the three corpora are varying in the recordings' length, ranging from short utterances containing single words in SUSAS, mid-sized sentences in eNTERFACE to long sentences in EmoDB, we had to apply different feature extraction methods to each dataset.

This was necessary to generate a sufficently large training set for each of the experiments. At the same time this provided an opportunity to look into possibilities for virtual expansion of given datasets.

Theoretically, CNNs are robust against different sizes and transformation operations (e.g. projection or rotation) of input patterns. But, unfortunately, they tend to overfit (cf. [27]) in the training process if not a sufficient amount of training material is available. Therefore, CNNs are usually applied to larger datasets. For example, the 2012 ILSVR Challenge consisted of over 10.000.000 samples labelled with roughly thousand classes (cf. [5]). In contrast, the EmoDB dataset consist of only 535 recordings split into seven classes. Additionally, there exist several methods to enlarge training sets of images quite easily, for instance, by mirroring or zooming of the relevant part. As this is not possible in spectrograms without changing the inherent meaning of the data, we had to employ a different approach.

In our experiments, we utilised the following approaches (applied to both training and test material) for the snippet generation depending on the utterances' length per corpus: For the SUSAS set we used the spectrogram calculated over the whole respective recording. For EmoDB and eNTERFACE we split the long sentences into parts of 1.5 sec. Additionally, for further increase of training material, for all snippets the corresponding spectrograms were generated with three different maximum frequencies, namely 7, 7.5, and 8 kHz respectively. With this approach, we bootstrapped the given material (cf. [28]) to achieve a larger amount of samples necessary for a CNN training and testing. In particular, the method simulated small differences in the pronunciations of each speaker. Therefore, a better representation of the emotion characteristics with respect to generalisation was achieved, especially in the Leave-One-Speaker-Out (LOSO) experiments. The number of original and bootstrapped input samples is given in Table 2. Since our approach only stretched the original spectrogram we did not expected any aliasing effects. This aspect was manually cross-checked. Furthermore, we tested this method with a simple CNN on EmoDB and obtained a (slight) improvement from 0.58 (no bootstrapping) to 0.61 (bootstrapped) unweighted recall.

4.2. Architecture

In the CNN-related community a few network architectures are widely used for classification (cf. Section 2). The network applied in our experiments was inspired by the original AlexNet (cf. [16]) and was adapted to the specific needs

[1] It is to be noticed that depending on the corpus the snippet's length varied.

TABLE 2. SIZE OF THE DATASETS AFTER EXPANDING RELATED TO THE VARIATION APPLIED TO THE CORPUS.

	Number of Input Data		Variation	
	Original	Bootstrapped	Time windows	Variable Frequency
EmoDB	535	17092	yes	yes
eNTERFACE	1277	19144	yes	yes
SUSAS	3593	10779	no	yes

of the classification task. In particular, the network is constructed as follows: Five convolutional layers are combined with two pooling layers. Their output is further processed in two activation layers employing Rectified Linear Units (ReLUs) as activation functions together with a respective dropout layer to reduce the effect of overfitting. Finally, fully-connected layers classified the emotional content of the speech utterances. The whole network is schematically presented in Figure 2 and the layer's parameters are shown in Table 3.

TABLE 3. MAIN LAYERS OF THE CNN-BASED ARCHITECTURE'S PARAMETERS.

Layer Name	Filter Size	Step Size	Depth
Input Layer			
Conv 1	11x11	5	96
Max Pooling	2x2	2	
Conv 2	5x5	1	256
Conv 3	3x3	1	384
Conv 4	3x3	1	384
Conv 5	3x3	1	256
Max Pooling	2x2	2	
ReLU Dropout Layer	4096		
ReLU Dropout	4096		
Output Layer	Variable		
Softmax			

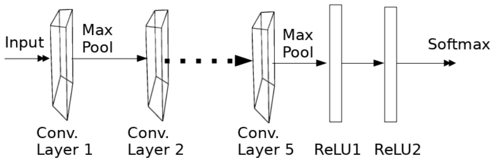

Figure 2. Schematic overview of the architecture combining five convolutional layers and two ReLUs. In Table 3 the layers' parameters are listed. The system is adapted from the AlexNet (cf. [16]).

As described in Section 4.1, we generated spectrograms from the audio material provided by the corpora. The size of each spectrogram was defined as 256x96 pixel resulting in the first kernel's size of 11x11 pixels. Using a step size of 5, we obtained 96 generated kernels in the first convolutional layer. In the next step, we applied a 2x2 max pooling layer with a step size of 2. This reduced the amount of generated input which was passed to the next convolutional layer about roughly 50%. Given this subnetwork, we derived a kind of feature extraction automatically processed by the system.

In the following we applied four additional convolutional layers to the preprocessed material where the corresponding parameters are given in Table 3. These layers allowed a step-by-step focussing on the inherent emotions coded in the audio samples. Again, a max pooling layer was inserted to further concentrate the provided information.

The final assignment of labels was done in the third subnetwork. For this, we chose ReLUs as activation functions since they are fast in calculation and further, robust against saturation. Furthermore, in the dropout layers we employed for the training a rate of 50% as this was shown to be an optimal value (cf. [27]). Finally, the output was realised with a softmax function.

4.3. Validation Measures

In our experiments, we applied the LOSO validation paradigm. For this, the network was trained on the samples derived from all speakers except one speaker. This material was used for testing only. In general, this strategy was applied to each speaker provided in the corpus. In addition to the test set we used a small part of the training set for development purpose deciding when to stop training.

The evaluation is based on the unweighted recall (UR) achieved in each LOSO run. These values are shown in Table 4.

In the description of our results we used the top notation which is sometimes provided in the community allowing a better interpretation of tendencies in the trained networks. The top1 value shows the achieved recall if only the class which is ranked highest by the network is considered. This is comparable to the standard recall values mainly used in classification tasks. The results mentioned in top2 and top3 extended the investigation towards the second and third best ratings. This means, the target class known from ground truth is within the labels assigned with the three highest ranks. For this, a trend in the classification could be better derived, especially in cases where overfitting could be possible.

5. Results and Discussion

5.1. Classification Results

At first, we have to consider that in recent work in CNN-based emotion recognition from speech usually EmoDB was used as single corpus. In our experiments, we extended the convolutional approach also to less acted (e.g. eNTERFACE) and close-to-real-life (e.g. SUSAS) speech material. In addition, the two corpora also provide more acoustic material which raised the opportunity to face the problem of overfitting. In the following we report on results based on the data prepared as in Section 4.1. In particular, for EmoDB results on non-bootstrapped material are described in [11] and therefore, we concentrated on preprocessed material showing that this improves the classification performance.

Considering the EmoDB results we saw that the classifier showed a clear tendency for overfitting on the development set since an average recall of 0.94 was achieved.

In contrast, the classification results on the test set showed values comparable to those reported in the literature. For instance in [25], a prediction accuracy averaged across all emotions of 0.52 and average across all speakers of 0.84 is achieved. It is to be noticed that the reported results were achieved in cross-validation experiments which have less requirements according to generalisation. In the LOSO paradigm the classifier needs a higher generalisation ability, but we already achieved an UR of 0.71 in the top1 case. Considering the three highest ranked ratings we obtained a performance of 0.94 on the test set (cf. Table 4). This showed that the proposed architecture can be used for emotion classification from speech even compared to other classification approaches (e.g. Hidden Markov Model (HMM) with UR 0.73 [8]).

TABLE 4. AVERAGED UNWEIGHTED RECALL (UR) FOR LOSO EXPERIMENTS ON EACH CORPUS, AVERAGED ACROSS ALL SPEAKERS PER CORPUS. BESIDES THE TOP1-3 VALUES ON THE TEST SET THE AVERAGE RECALL ON THE DEVELOPMENT SET IS SHOWN.

	Avg. UR on test set			Avg. UR on development set
	top1	top2	top3	top1
EmoDB	0.71	0.86	0.94	0.96
eNTERFACE	0.66	0.73	0.86	0.87
SUSAS	0.57	0.76	0.89	0.92

Based on the results of EmoDB, we extended our investigations towards corpora with lower recording qualities.
In Table 4) the UR results on SUSAS, a corpus providing spontaneous emotional speech, are shown. Given the lower quality in recording and the less expressive emotions the recognition performance is still comparable to published results. In [8] a UR of 0.55 on average is reported on SUSAS using HMMs whereas with CNNs an average UR of 0.57 was achieved (cf. Table 4). As it can be seen from the development set the overfitting effect is minimised due to the larger amount of available acoustic material.
Considering the eNTERFACE corpus, using the same paradigm as on EmoDB and SUSAS we achieved an average UR of 0.66 (cf. Table 4). Though this corpus provides emotional speech which is acted the classification performance is lower than expected (cf. the performance on EmoDB). From our point of view, this is mainly related to the huge diversity in the speech. As mentioned in [9] all speakers speak English although most of them are non-native speakers. Further, a large variety of accents and dialects are present in the corpus. Therefore, the automatic extraction of reasonable parts from the spectrograms done by the CNN layers was complex. Nevertheless, the achieved average UR is comparable to other approaches (cf. UR of 0.67 in [8]).

5.2. Deep Dreaming

Using the deep dreaming approach proposed by Google (cf. [13]) it is possible to visualise which features applied for the network's training seem to be important. This is usually done by superimposing the input's patterns with information supporting the target class on output. For this, a highlighting effect is generated showing the most likely parts of an image related to a certain class. This means that a network trained for emotion recognition from speech (i.e. on spectrograms) tries to enhance the parts in the input image which resemble at least one of the target emotions. In our experiments, we applied deep dreaming also for an interpretation of the internal representation of speech related features.

In the following we discuss sketchy results of the deep dreaming approach: Applying deep dreaming to a spectrogram showed in which way different parts of the image have been highlighted during several iterations of the method. In Figure 3 the original input is shown on the left. Additionally, five iterations of deep dreaming are presented. Comparing the input spectrogram and the first iteration we see differences in the heatmap representing the amplitudes. In the first iteration's heatmap most amplitudes are emphasised (more intense red and yellow in the spectrogram) since in the beginning most spectrogram's parts seem to be important for classification. We conclude that the network used this particular information in conjunction with the relation of spectral peaks as an indicator for the emotions.

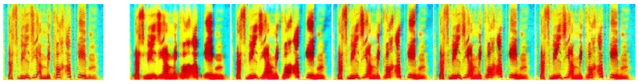

Figure 3. Example of a spectrogram taken from EmoDB processed by deep dreaming. On the left, the original input is given while the images on the right visualise a sequence of spectrograms iteratively processed. The colours' intensity is related to the highlighted parts of the image.

Further iterations of deep dreaming lead to an increase of details in the generated visualisation focussing on particular amplitudes related to classification. Additionally, during the learning process the original input will be stepwise recreated. On the one hand, this effect can be used for inputs afflicted with noise, on the other hand it shows also that the CNN-based network can tend to overfitting. In Figure 3 we see that the rightmost image is quite similar to the input (leftmost spectrogram) since just a few amplitudes are emphasised while in the first iteration the network is still sensitive to most of the input parts. Nevertheless, some parts of the spectrogram have been highlighted (more intense colours) indicating that these features are still reasonable for emotions classification also in a fully-trained network.

Given these indications, we can assume that in future a more detailed analysis of features represented in the spectrogram and highlighted by deep dreaming can support and influence the still ongoing discussion on suitable feature sets for speech emotion recognition.

6. Conclusion

In the current paper, we presented experimental investigations on CNNs in the emotion recognition from speech. For this, the audio material was transferred to spectrograms which can be used for CNNs. In particular, we proposed a neural architecture based on the AlexNet (cf. [16]) which shows comparable results to other approaches discussed in the literature. Furthermore, we also discussed a method to increase the amount of training material necessary for a proper training of such networks. The presented methods reasonably increase the training material while simultaneously information was neither lost nor highly corrupted. In contrast to other works related to emotion classification with CNNs (cf. [24], [25]), we applied our approach to three corpora, namely EmoDB, eNTERFACE, and SUSAS, demonstrating that the architecture can handle material generated in various conditions.

Moreover, the internal representation of features extracted by the CNN architecture was investigated using the deep dreaming method (cf. [13]). For this, we found areas in the spectrogram which suggest spectral parts related to certain emotions. Besides further optimisation of the CNNs' parameters, these feature related analyses will be part of future investigations. This is highly linked to the discussion on optimal feature sets. Additionally, we suggest to apply deep dreaming not only for feature inspections but also for visualising emotions classes like anger, joy, or valence/arousal to derive conclusions for optimal training data in terms of features and prototypical acoustics.

Acknowledgments

We acknowledge continued support by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation and by the project "Mova3D" (grant: 03ZZ0431) sponsored by the German Federal Ministry of Education and Research.

References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proc. of 2013 Intern. Conf. on Acoustics, Speech & Signal Processing*. Vancouver, Canada: IEEE, 2013, pp. 8599–8603.

[3] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. of 2011 Intern. Conf. on Acoustics, Speech & Signal Processing*. Prague, Czech Republic: IEEE, 2011, pp. 5688–5691.

[4] E. M. Albornoz, M. Sánchez-Gutiérrez, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, *Spoken Emotion Recognition Using Deep Learning*. Springer, 2014, pp. 104–111.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Intern. Journ. of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[6] K.-C. Wang, "The feature extraction based on texture image information for emotion sensing in speech," *Sensors*, vol. 14, no. 9, pp. 16 692–16 714, 2014.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. of INTERSPEECH-2005*. Lisbon, Portugal: ISCA, 2005, pp. 1517–1520.

[8] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. of the ASRU 2009*. Merano, Italy: IEEE, 2009, pp. 552–557.

[9] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. of the Workshop on Multimedia Database Management*. Atlanta, USA: IEEE, 2006, s.p.

[10] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. of EUROSPEECH-1997*. Rhodes, Greece: ISCA, 1997, pp. 1743–1746.

[11] N. Weißkirchen, "Implementation von convolutional neural networks und deren anwendung in der emotionserkennung aus sprache," Master's thesis, Otto von Guericke University Magdeburg, 2017.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps." *CoRR*, vol. abs/1312.6034, 2013.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of 2015 Conf. on Computer Vision and Pattern Recognition*. Boston, USA: IEEE, 2015, pp. 1–9.

[14] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. on Audio, Speech, & Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Intelligent signal processing*. IEEE Press, 2001, pp. 306–351.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*. Curran Associates, Inc., 2012, pp. 1097–1105.

[17] M. D. Zeiler and R. Fergus, *Visualizing and Understanding Convolutional Networks*. Zurich, Switzerland: Springer, 2014, pp. 818–833.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of Computer Vision and Pattern Recognition*. Las Vegas, USA: IEEE, 2016, pp. 770–778.

[19] M. Kleinschmidt, "Robust speech recognition based on spectro-temporal processing," Ph.D. dissertation, University Oldenburg, 2002.

[20] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proc. of the 22nd ACM Intern. Conf. on Multimedia*. Orlando, USA: ACM, 2014, pp. 801–804.

[21] D. Palaz, M. Magimai.-Doss, and R. Collobert, "Analysis of cnn-based speech recognition system using raw speech as input," in *Proc. of the Interspeech*. Dresden: ISCA, Sep. 2015, pp. 11–15.

[22] B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Computer Speech & Language*, vol. 28, no. 2, pp. 483 – 500, 2014.

[23] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, pp. 1–23, 2012.

[24] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. of 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf.* Jeju, South Korea: IEEE, 2016, pp. 1–4.

[25] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. of 2017 Intern. Conf. on Platform Technology and Service*. Busan, South Korea: IEEE, Feb 2017, pp. 1–5.

[26] J. Smith, *Mathematics of the Discrete Fourier Transform (DFT)*. W3K Publishing, 2007.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.