Emotions in Spoken Language - Do we need acoustics?

Nadine Probol
University of Applied Sciences
Darmstadt
nadine.probol@h-da.de

Margot Mieskes
University of Applied Sciences
Darmstadt
margot.mieskes@h-da.de

Abstract

Work on emotion detection is often focused on textual data from i.e. Social Media. If multimodal data (i.e. speech) is analysed, the focus again is often placed on the transcription. This paper takes a closer look at how crucial acoustic information actually is for the recognition of emotions from multimodal data. To this end we use the IEMOCAP data, which is one of the larger data sets that provides transcriptions, audio recordings and manual emotion categorization. We build models for emotion classification using text-only, acoustics-only and combining both modalities in order to examine the influence of the various modalities on the final categorization. Our results indicate that using text-only models outperform acoustics-only models. But combining text-only and acoustic-only models improves the results. Additionally, we perform a qualitative analysis and find that a range of misclassifications are due to factors not related to the model, but to the data such as, recording quality, a challenging classification task and misclassifications that are unsurprising for humans.

1 Introduction

The correct detection of emotions in spoken language is a task which has been examined a lot in recent years. But the majority of research so far, has treated text and speech separately and there is little research on using both audio and textual data for emotion recognition, which would stand to reason as emotions are expressed not only in *what* is being said, but also *how*. One exception is Ho et al. (2020) who present a model using Multi-Level Multi-Head Fusion Attention mechanism and recurrent neural network (RNN) for the detection of the audio input in combination with information gained from text.

Building on previous work, we show extensive experiments on various types of models trained on text as well as speech. We examine, whether the combination of emotion recognition from textual data as well as audio data improves the results to either singular approach. We also explore different feature sets and models to find the best combination. This allows us to conclude, that, while using only textual data yields reasonable results for English, we achieve best results when combining it with additional acoustic data.

In a qualitative analysis we take a look at the most striking confusions between classes and give possible reasons for them. Additionally, we give an idea of how to counter these problems in future experiments.

In Section 2 we describe previously conducted work on emotion recognition. For this, in Section 2.1 we describe some of the more often used data sets available for this task. In Section 2.2 we describe some related work in more detail. The experimental setup is described in Section 3. It additionally focuses on the feature extraction for the audio data (Section 3.1), the preprocessing of the textual data (Section 3.2, as well as the model used for the transcriptions (Section 3.3 and the method of combination for the different models (Section 3.5. Section 4 shows the best results of the combined models and three different data set variations. In Section 5 we conduct a qualitative analysis of the results. Finally, we conclude in Section 6.

Our major contributions therefore are:

- A comparison of results on text only, speech only and the combination thereof.
- An analysis of various combinations of acoustic and textual features and models for the classification of emotions from speech data.
- A qualitative analysis of the problematic cases and an investigation into the sources of these cases.

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 71–84
July 14, 2023 ©2023 Association for Computational Linguistics

2 Previous Work

The detection of emotions in spoken language is a complex problem. Emotions in humans are expressed both in the choice of words used, but also in the way these words are expressed acoustically. The related work on emotion detection based on transcribed speech, acoustical information and a combination thereof, described below is focused on relevant experiments related to our work and does not represent an exhaustive review of the topic.

2.1 Emotion Data sets

There are various data sets for multimodal emotion detection, such as IEMOCAP[1], EmoDB,[2] DES,[3] SAVEE,[4] or CASIA.[5] These data sets cover various languages such as English (IEMOCAP), German (EmoDB), Chinese (CASIA, NNIME[6]) or Danish (DES).

Out of these, the IEMOCAP data set is one of the larger ones. Busso et al. (2008) introduced the IEMOCAP (Interactive Emotional dyadic Motion Capture Database) data set. It contains 9,924 example utterances and corresponding labels, spoken by ten different actors (five male, five female), resulting in a quite balanced set with a slightly higher amount of female data (51.37%, Figure 5 in Appendix C). The most common length for an utterance is one word (701 utterances), followed by six words (668) and five words (653) and the examples are further divided into five sessions, containing 80 improvised and 70 scripted dialogues (Figure 6 in Appendix C). Overall, the data set contains slightly more improvised examples (52.95%) than scripted ones (47.05%), but all recordings are of actors, so there are no naturally occurring dialogues. In the proposed models, each labelled example will be used without the context of the given dialogue and corresponding session. Each utterance has been annotated by three different annotators (out of six) and labelled as one of ten categories (Anger, Happiness, Sadness, Neutral state, Frustrated, Excited, Fear, Surprise, Disgust, Other), which differ greatly in size (Figure 7 in Appendix C).

The data set includes the full transcripts, hand-annotated emotion labels and the audio recordings.

The IEMOCAP data set has been widely used for example by Mirsamadi et al. (2017), Mao et al. (2019), Dangol et al. (2020), or Lieskovska et al. (2022).

2.2 Methods for emotion detection

Emotion recognition of speech is either done text-only or acoustics-only, combined models are rarer. In the following, we present work from all three areas as examples.

Acoustic data

Based on their work, Lieskovská et al. (2021) concluded, that the usage of deep convolution architectures, which are based on spectrum information only is increasing. The authors considered these architectures as well as recurrent networks as a strong base for emotion recognition systems for speech. They state that, even though many used attention mechanisms to improve the performance of their model, the magnitude of improvement is unclear, which makes this approach dispensable.

An interesting experiment was conducted by Mirsamadi et al. (2017), who compared a neural network with an SVM on the emotions happy, sad, neutral, and angry of the IEMOCAP data set. The authors proposed a deep neural network with two hidden layers followed by an LSTM layer. They used emotion LLDs with RNN-weighted pooling with a logistic regression attention model. In their work, this approach performed best, by focusing on emotional parts of utterances. They also trained an SVM on mean, standard deviation, minimum, maximum, range, skewness and kurtosis. Overall, the authors concluded, that the SVM approach needed a higher amount of statistical functions to reach its best performance, whereas the DNN was less sensitive to number and diversity of the used statistical functions.

Textual data

Also, there are interesting experiments for the research on textual data. For example, Mohammadi et al. (2019) compared the results of SVMs, neural networks and a combination of the two. The authors used the pre-trained ELMo word embedder by Peters et al. (2018). Following the input layer were two layers of 25 bidirectional GRUs and an attention layer. These steps were done three times in parallel and their output was then concate-

[1] https://sail.usc.edu/iemocap/

[2] http://emodb.bilderbar.info/docu/

[3] http://universal.elra.info/product_info.php?cPath=37_39&products_id=78

[4] https://www.tensorflow.org/datasets/catalog/savee

[5] http://www.chineseldc.org/resource_info.php?rid=76

[6] https://nnime.ee.nthu.edu.tw/

nated. Additionally, the authors used an SVM with polynomial kernel with a degree of 4 and set C to 2.5. Using a neural network for feature extraction and an SVM for classification gave the best results. However, as it was for task 3 at SemEval 2019[7], they only classified three classes: angry, happy, and sad.

Chakravartula and Indurthi (2019) present a model with a stacked BiLSTM architecture for the SemEval 2019 task 3, which is based on written dialogues. The authors used three different embedding layers: The first embedding layer converts each word into its corresponding 300 dimensional GloveEmb word vector, the second takes the POS tags and converts each of them into a constant one-hot vector and the third embedding layer converts each word into a vector based on the values in the DepecheMood affective lexicon (Staiano and Guerini, 2014). They achieved the best results by combining the first and third embedding layer with two BiLSTM layers, however, combining the first embedding layer with two BiLSTM layers and attention reached comparable results.

Acoustic and textual data combined

There is little research on combining acoustic and textual data for emotion detection. The following are the most important for the work at hand.

Yoon et al. (2018) built two encoders: The Audio Recurrent Encoder (ARE) and the Text Recurrent Encoder (TRE), which work in parallel. For the audio encoder, they use MFCCs and prosodic features, which they extract via the *openSMILE* toolkit. By using the *NLTK* toolkit, the authors tokenized and indexed the transcripts into a sequence of tokens. Both, the ARE and TRE use RNNs to each predict an emotion class. For a final prediction of both models together, the authors use a softmax function to concatenate the vectors of the predictions of the audio RNN and text RNN. Later, the authors improved their model (Yoon et al., 2019) by using a bidirectional encoder (BRE) for both the textual and audio data instead of unidirectional. The final hidden representation of the audio-BRE is then used as a context vector to bring attention to the text-based vector. The authors then apply second-hop attention with this information to the audio sequence, which they call MHA-2. Both times the authors used a variation of the IEMOCAP data set (using only the categories angry, happy, sad and neutral) to test their model.

Another approach by Ho et al. (2020) proposes a multi-level multi-head fusion attention model (MMFA). For the audio data, the authors extract MFCC features via the *openSMILE* toolkit as well. The audio frame is set to 100ms at a rate of 50ms and a Hamming window is applied, so the temporal length of the audio is ten times longer than the length of the utterance pre re-sampling. The authors state, that the attention mechanism in MMFA combines the contextual information of audio and text. The final model can be divided into two parts: first-level attention and second-level attention. The first part computes a representation for each audio and text RNN-feature at different positions of a single sequence. The second attention is a modified multi-head attention, in order to fuse the attention features from both audio and text. It does not compute this attention just once, but multiple times in parallel. The authors use two different versions of the IEMOCAP data set to test their models: one with four classes ("Neutral state", "Anger", "Sadness", and "Happiness/Excited") and one with eight classes (the original classes of the data set minus "Disgust" and "Other"). Also, they look at two scenarios: improvised (using only the improvised examples of the data) and mixed (using all data). When concentrating on the variation with four classes, the model achieves an accuracy of 0.73 on the mixed data and 0.77 on the improvised data. For the version with eight classes, the model reaches an accuracy of 0.57 for the mixed data and 0.61 on the improvised data.

3 Experimental Setup

For our experiments[8] we use the IEMOCAP data set by Busso et al. (2008). As, out of the above presented data sets (Section 2.1), the IEMOCAP data set is the largest and provides both textual and audio data in English. The EmoDB, CASI, NNIME and DES also provide both textual and audio data, but they are either very small (i.e. EmoDB only has 500 sentences) or in Chinese (CASIA, NNIME), which is a tonal language, as opposed to English, which is a non-tonal language and therefore, expresses emotions differently.[9]

[7] As only Task 3 of the SemEval 2019 workshop focuses on emotion detection (EmoContext), we only looked at the papers for this Task.

[8] All experiments are conducted on Windows 10 with Python 3.8.10. The additionally used libraries are listed with their corresponding version in Table 3 in Appendix A.

[9] Additionally, none of the authors are proficient in Chinese, which makes the qualitative analysis impossible.

The authors performed the emotion category annotation and report a Fleiss' Kappa of 0.27 on the entire annotation. Our re-calculation of Fleiss' Kappa resulted in an observed agreement of 0.23, an expected agreement of 0.27 and an overall agreement of -0.06.[10] However, Fleiss' Kappa is not applicable to data sets with empty annotations, which is the case in this data set, as there are six annotators in total, but only three given annotations per example. Thus, we use Krippendorff's Alpha to verify their results. This leads to an observed disagreement of 0.77, an expected disagreement of 0.73 and an agreement of -0.06, which confirms the low agreement between the annotators.

As the original data set does not provide a gold standard, we use a majority vote between the three annotations. If all annotators decide on different labels, we use the label of the first annotator. Another option would have been to randomly take one annotation, but by always deciding on the first annotator in these cases, it might lead to a slightly higher consistency. Table 1 shows the spreading of the distribution of classes in the different variations.

Emotion	Original	Variation 1	Variation 2
Anger	x	xX	xZ
Happiness	x	xY	xY
Sadness	x	x	xZ
Neutral state	x	–	x
Frustration	x	xX	xZ
Excited	x	xY	xY
Fear	x	–	–
Surprise	x	–	–
Disgust	x	–	–
Other	x	–	–

Table 1: Different variations of the IEMOCAP data set. Added capital letters show a combination of two or more classes. For example, In Variation 1, classes "Anger" and "Frustration" are combined (shown with a X), as well as the classes "Happiness" and "Excited" (shown with a Y).

The classes "Fear" and "Neutral state" are excluded in Variation 1. "Fear" is an extremely small class, though not as small as "Disgust". The class "Neutral state" is also excluded. Although it is the largest class, it shows no explicit emotion.

Variation 2 combines the classes "Anger", "Frustration" and "Sadness", as they are all negative emotions. It contains the combined class of the happy emotions of "Happiness" and "Excited" as well as the class "Neutral state". This class can be seen as a sentiment classifications (positive, negative and neutral).

Previous work show that some authors group different classes together based on similar acoustic signals. For example, Nwe et al. (2003) combined the classes "Anger", "Surprise" and "Joy" in one cluster and the classes "Fear", "Sadness" and "Disgust" in another one. This approach is based on findings by Williams and Stevens (1981) (cited by Nwe et al. (2003)), who found that emotions such as "Anger" and Fear" but also "Joy" arouse the sympathetic nervous system, while emotions such as "Sadness" arouse the parasympathetic nervous system. An aroused sympathetic nervous system leads to an accelerated heart rate and higher blood pressure, a dry mouth and even occasional muscle tremors. This shows in a loud, fast and enunciated speech with strong high frequency energy. In contrast, an aroused parasympathetic nervous system leads to lower blood pressure and heart rate, as well as an increased salivation. Speech produced under these circumstances is slow with little high frequency energy. Even though Nwe et al. (2003) improved their results by grouping the emotion classes accordingly (accuracy reaches up to 90%), it is questionable whether it is useful to the actual use case to subsume such different classes.

In our experiments, we combine emotions which are similar, such as "Happiness" and "Excited", or show the same sentiment ("Anger", "Frustration" and "Sadness" are all negative).

3.1 Feature extraction for audio data

To extract features from the audio files, we use two different approaches.

First, we use the *openSMILE* toolkit by Eyben et al. (2010), which is accessible via a Python API.[11] This API has six different extractable feature sets and we chose "ComParE_2016", which was first introduced for the *Interspeech 2016* (Schuller et al., 2016), as this provides the largest amount of different features extracted.

We extract features as two-dimensional tables along the time-axis. For each feature, we calculate the maximum, minimum, mean and standard deviation, which results in 100 different features. This is comparable to Mirsamadi et al. (2017), who also used the mean, standard deviation, minimum, maximum, and other features to train their models.

[10] Calculated using https://dkpro.github.io/dkpro-statistics/

[11] https://github.com/audeering/opensmile-python/

The second approach is based on the *librosa* library (McFee et al., 2015), which extracts Mel Frequency Cepstral Coefficients (MFCCs). The audio signal is split evenly into $slices$ of 10 ms and for each slice we extract 13 MFCCs. For longer recordings, this *can* result in files with more than $1,000$ slices. As this is the case for only 12 examples: Six from class "Sadness", three from class "Frustration", two from class "Excited" and one example from class "Anger", but considerably affects computing time, we cut off all MFCCs after $1,000$slices.

3.2 Preprocessing text data

We tokenize the transcription using the *NLTK* tokenizer.[12] Then, we used part of the code from the *tensorflow* tutorial for word2vec[13] to change those tokens to numbers, so the model can process them, as well as add them to a vocabulary dictionary. The maximum word vector length is 200. Longer examples are cut off after the 200th word and shorter examples are filled up with zeros. Zeros have no word associated within the vocabulary dictionary as they are used as padding. Additionally, we change the class labels to numbers and then change those numbers to categorical tensors.

3.3 Model for Transcripts

For the transcriptions, we trained a neural network with bidirectional LSTM and GRU layers and set the vocabulary size to 15,000. Bidirectional LSTM layers are used in many other experiments, such as the ones by Chakravartula and Indurthi (2019) and seem to achieve good results. There are not as many experiments using GRU layers, however, the aforementioned experiment by Mohammadi et al. (2019) achieved good results using GRU layers. Chatterjee et al. (2019) described that most participants in the SemEval 2019 Task 3 were using LSTM and BiLSTM models, though GRU and CNN models were also used by a few teams. The complete model, including hyperparameters, we use for the transcripts is visualized in Appendix B (Figure 4).

3.4 Model for Acoustics

In general, we trained SVMs and neural networks on the acoustic data.

First, we trained an SVM on all standard derivations (named SVM SD in the following) and one on all 100 available features extracted through *openSMILE* (SVM 100).[14]

Second, we use the MFCCs extracted through the *librosa* library to train various neural networks, as recurrent networks seem to be a good basis for emotion recognition systems (Lieskovská et al., 2021). This is comparable to Wang et al. (2015), however, we focused on the statistical parameters and used a different data set. Based on preliminary experiments done by ourselves, as well as on experiments of Mirsamadi et al. (2017) and Dangol et al. (2020), we test different architectures with LSTM and GRU layers and different combinations to improve the results. Additionally, we test the different usages of ReLU or SELU activation functions. By using SELU activation we avoid running into the so called dying ReLU problem, which was described for example by Agarap (2018) or Lu et al. (2019). The dying ReLu problem refers to neurons becoming inactive and therefore only have output 0 for any input. SELU activation function induces self-normalization, which is faster than an external normalization, and therefore leads to a faster convergence of the network. Contrary to ReLU, it can go below 0, avoiding dying neurons.

The exact models and their names are listed in Table 5 in Appendix D.

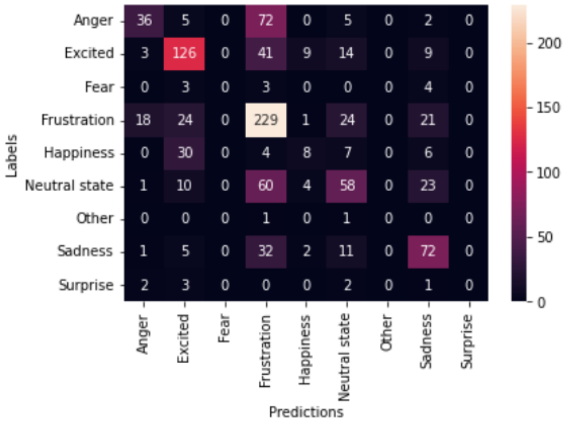
One model (Audio NN 4 as named in Table 5 in the Appendix) is more complex than the other models as it consists of a convolutional 2D layer with ReLU activation and 32 hidden units, followed by a max pooling 2D layer and a batch normalization layer. These layers are then following once again. After, a dense layer with 64 hidden units, SELU activation and L2 regularization follows and again a max pooling 2D and batch normalization layer. Then comes a flatten layer and another dense layer with 64 hidden units and ReLU activation. After a dropout of 0.3 follows the same output layer as in the former models.

For the *openSMILE* features, we used similar architectures, as for the *librosa* features. Differences in the architectures are due to the different shape of the data, which for example results in the usage of a standard input layer as an input for the NNs trained on *openSMILE* features, whereas this is not possible for *librosa* features.

[12] https://www.nltk.org/_modules/nltk/tokenize.html

[13] https://www.tensorflow.org/tutorials/text/word2vec

[14] Additionally trained SVMs are listed in Table 6 in Appendix D.



	Anger	Excited	Fear	Frustration	Happiness	Neutral state	Other	Sadness	Surprise
Anger	36	5	0	72	0	5	0	2	0
Excited	3	126	0	41	9	14	0	9	0
Fear	0	3	0	3	0	0	0	4	0
Frustration	18	24	0	229	1	24	0	21	0
Happiness	0	30	0	4	8	7	0	6	0
Neutral state	1	10	0	60	4	58	0	23	0
Other	0	0	0	1	0	1	0	0	0
Sadness	1	5	0	32	2	11	0	72	0
Surprise	2	3	0	0	0	2	0	1	0

Figure 1: Confusion Matrix of the combination of the Text NN, SVM SD and SVM 100 on the original data set

The exact specifications of each model are listed in Table 4 in Appendix D and link them to the names we use for the single models.

3.5 Method of combination

In order to combine results from the acoustical classification model and the textual classification model, we use two different approaches:

The first approach combines every model with every other model with a weighting of 50% to 50%. Yoon et al. (2018) and Yoon et al. (2019) combine their models with this weighting, however, they always combine their models trained on textual features with one trained on acoustic features. The combinations in our experiments can also include only models based on acoustic features, as the ablation process only focuses on the best results and not the feature type.

The second approach of combining models is very similar to the first one. Here, we combine three models with a weighing of 33% to 33% to 33%.This also excludes the combination of two or three same models, so there is no single model prediction or a weighing of 33% to 67%. With this method of combination, 21,168 different combinations are possible which we all tested and compared their results in order to find the best combination.

4 Results

To classify the results, we compare them to a majority baseline (Table 2).

4.1 Original data set

Table 2 shows the best results for the original data set classifying all emotions available. The best single model is the Text NN. By combining the model with the SVM 100, the results improve, however, combining those two models with the SVM trained on the standard derivations of all features gives the best results on the original data set (Section 3.4).

This leads to an accuracy of 0.53, which is much lower than Ho et al. (2020), however, they dropped the smallest classes "Disgust" and "Other", which do have a negative impact on our model performance.

A look at the confusion matrix in Figure 1 shows, there are three classes, that do not get predicted at all. The combined models do not predict any examples as "Fear", "Other" or "Surprise". This aligns with class size, as these three classes are much smaller than the other classes (except for "Disgust"). This also applies to class "Disgust", which is not visible in the confusion matrix, as there is no example of it in the test split. There are only two examples of class "Disgust" in the data set, which were both automatically sorted into the train split.

Only 33.3% of all examples predicted as "Happiness" are correctly identified as such. Most confusions happen on class "Frustration". 48.2% of the examples predicted as "Frustration" do not belong to this class. 33.8% of these misclassifications are examples of class "Anger", which also means, there are more examples of class "Anger" predicted as "Frustration" (60.0%), than correctly identified. These confusions are also the most likely ones with these combined models (15.5% of all misclassifications). The second highest amount of misclassifications happen with examples of class "Happiness" as "Frustration" (12.9% of all misclassifications). There are also slightly more confusions of examples of class "Happiness" as "Frustration" (38.5%), than correctly identified (37.2%).

A more detailed analysis based on the single classes of the data set can be found in Section E in the Appendix.

4.2 Data set variation 1

Table 2 shows the best results for data set variation 1 which combined "Anger" with "Frustration", "Happiness" with "Excited" and kept "Sadness" and "Surprise" as separate categories. The results on data set variation 1, as for the original data set,

Variation	Models	Baseline	Macro Precision	Macro Recall	Accuracy	Macro F1
Original	Text NN	0.32	0.30	0.28	0.48	0.28
	Text NN + SVM 100		0.32	0.29	0.50	0.30
	Text NN + SVM SD + SVM 100		**0.34**	**0.31**	**0.53**	**0.31**
Variation 1	Text NN	0.49	**0.74**	0.52	0.70	0.54
	Text NN + SVM 100		0.53	0.52	0.70	0.52
	Text NN + Audio NN 4 + SVM 100		0.58	**0.54**	**0.76**	**0.55**
Variation 2	Text NN	0.55	0.62	**0.59**	0.69	0.59
	Text NN + SVM 100		0.65	**0.59**	0.71	0.59
	Text NN + SVM SD + SVM 100		**0.72**	**0.59**	**0.72**	**0.60**

Table 2: Results on different data set variations and their baselines. The best results are marked in bold.

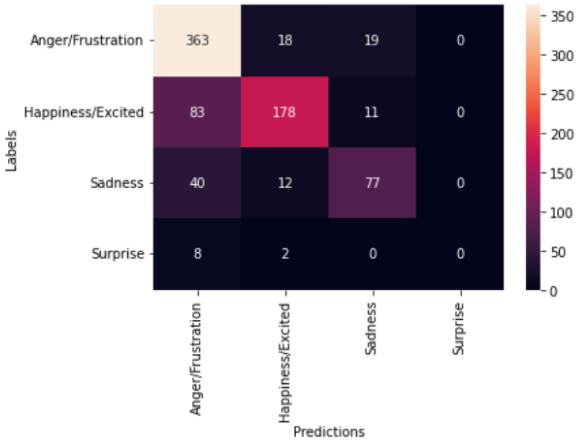


Figure 2: Confusion Matrix of the combination of the Text NN, Audio NN 4 and SVM 100 on data set variation 1

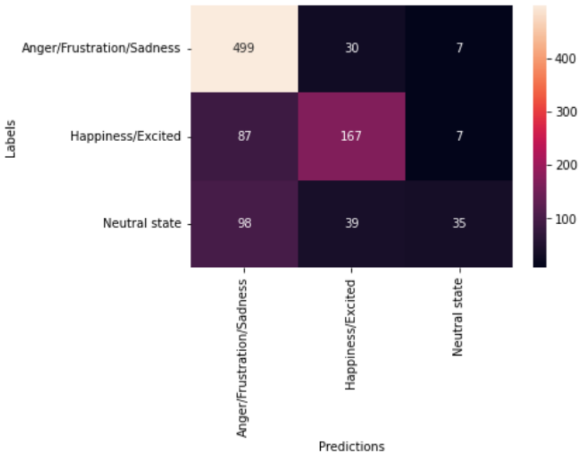


Figure 3: Confusion Matrix of the combination of the Text NN, SVM SD and SVM 100 on data set variation 2

show the best single model as the Text NN. However, the best combination of two models (Text NN and SVM 100) reach lower results. The combination of three models reaches also the best results out of all our experiments by combining the Text NN with the Audio NN 4 and SVM 100.

This combination achieves an accuracy of 0.76, which is higher than the accuracy Ho et al. (2020) reach on their four classes model (0.73). However, the classes are slightly different, as they detected neutral, angry, sad, and happy/excited. Further information on the performance of the single classes can be found in the Appendix E.

Overall, most correct classifications happen on class "Anger/Frustration" (58.7% and 50.9% of all correct predictions).

On the combination of the Text NN with Audio NN 4 and SVM 31 (Figure 2), the class "Happiness/Excited" is most often confused with other classes. 30.5% of the examples of class "Happiness/Excited" are predicted as "Anger/Frustration", which make up 43.0% of all misclassifications. The second highest amount of misclassifications happen with examples of class "Sadness" as "Anger/Frustration" (20.7%). The examples of class "Surprise" are mostly predicted as "Anger/Frustration", even though there are only ten examples on the test split.

4.3 Data set variation 2

The best results on data set variation 2 (which combines "Anger", "Sadness" and "Frustration", "Happiness" and "Excited" and keeps "Neutral state" as a separate category) are also shown in Table 2. The best single model is, like on the former data set variations, the Text NN, whose results improve by adding SVM 100. Again, the combination of Text NN, with SVM SD and SVM 100 (similar to the model combination for the original data set Section 4.1) achieves the best results (Accuracy of 0.72). While the recall remains the same for the single, two and three combined models, the additional information from the other models does specifically improve precision. More details on the single classes can be found in Appendix E.

The confusions matrix (Figure 3) shows 57.0% of the examples of class "Neutral state" are predicted as "Anger/Frustration/Sadness". This also represents the highest amount of confusions between two single classes (36.6% of all misclassifications). The second highest amount of misclassifications happen with examples of class "Happiness/Excited" as "Anger/Frustration/Sadness" (32.5% of all misclassifications).

Most correct predictions are on class "Anger/Frustration/Sadness" (71.2% of all correct predictions).

5 Qualitative Analysis

In our qualitative analysis, we take a closer look at those examples that have been misclassified. In general, most of the examples which are miscategorized are very short.

Also, we observed that there are examples that are not correctly identified by neither of the three models and data set variations. A closer look, revealed they are spoken by a female, however, in some instances a male speaker can be heard clearly in the background. While the female seems to be very happy in those, the background noise of the male could have negatively impacted the recognition process to "Anger" as he sounds quite angry.

Also, there was an example, which was predicted as "Excited", even though it was meant to be "Sad" as indicated by the manual annotation. This was interesting, as the person speaking sounded quite desperate, which shows more energy than the average sad person, which normally exhibits a rather low energy level. Therefore, the wrong classification as "Excited" makes sense and should be addressed in further experiments.

Additionally, we looked at some examples, which are wrongly classified in at least two of the three models.

After a proper examination, there are two types of confusions which stand out in particular: **Negative emotions which are confused as happy ones**. The aforementioned confusion of an example of a desperate person was not happening in just one example. Additionally, there was an example of a frustrated male which was quite energetic. However, in the background a woman was starting to speak in a higher voice, but the recording was cut shortly thereafter, making it impossible to understand what the woman ways saying. Her higher voice might have influenced the levels of the acoustic statistics we trained our model on.

Happy emotions which are confused as negative ones. Those happy examples tend to present low energy, for example, a person states "I love you a great deal!". This is a very happy statement, however, as the person sounds very close to tears, the model predicts it as "Frustration". Another person is very calmly speaking, however, due to the calmness, the person appears to the model to be sad. Both misclassifications do make sense as, as described by Nwe et al. (2003), speech representing sadness is slower and characterized by less high frequency energy.

Another noteworthy aspect is the length of the examples. While Seehapoch and Wongthanavasu (2013) already concluded that it is more difficult to correctly recognise the emotion if the speech is too long, our findings suggest the same is true for examples which are too short. There are several very short examples on the data set (approx. 1s), which are incorrectly classified. While some of them do express emotions, it can be hard even for humans to correctly determine them without context.

Many of the wrongly predicted examples also could be categorized in various classes, which explains the low inter-annotator-agreement.

6 Conclusion

In our work, we examined the benefit of combining information from text with information from speech in order to categorize emotions in spoken language. To this end, we used the IEMOCAP data set, in its original classification, but also in different combinations, to train various machine learning models (SVM and neural networks) to classify emotions. We combined the trained models in various ways to find the best combination of models for the classification.

Our results indicate that a combination of text-based features together with acoustic features provide the best results, as all combinations contained both models based on textual information, as well as acoustic information.

A detailed look at the results reveals that neural networks trained on textual data perform best, when only one modality (text) is used. However, when multimodal data (text and speech) is available, making use of all modalities improves the classification and that textual data is crucial for a successful classification. Our results also indicate, that using only acoustical data gives results that are even below those for text-only based classification. So yes, we need acoustics, but acoustics alone do not provide enough information for a successful classification.

Also, there is no clear indication, if *openSMILe* or *librosa* features provide better results. But, we observed that the SVMs trained on various acoustic features perform much better than the neural networks trained on the same features. This is probably due to the relatively small data set size.

In general, it seems that negative classes are easier to classify than positive emotions. As our results are only based on one data set, it would be worthwhile to explore this in more detail.

Additionally, we see a systematic problem with the correct classification of examples with unusual energy levels for the respective emotion. Therefore, it might be an interesting approach for examples of the class "Excited" to be divided into positive and negative excitedness. However, this would also need larger data sets, as the class "Excited" is already quite small in the IEMOCAP data set, which is already one of the larger ones available. To avoid the confusion of happy expression with unusually low energy levels, it might be important to add more examples like these to the training data. A first approach to do so could be achieved by simply oversampling happy examples with statistically low energy levels.

In general, some of the examples of the data set are hard to understand even for humans. This is due to poor audio quality as well as background noise. Sometimes we needed to look at the label in order to decipher which talking person was labelled, which also accounts for the low inter-annotator agreement we observed. It might be a useful approach to detect these examples and drop them before training a model, as this does only apply to a small part of the data set.

Broadly speaking, a combination of more audio features improves the results in comparison to single audio features. However, it should be further investigated, whether focusing on specific features might improve the results even more. A focus hereby should lay on the standard deviation of the features, as they reach the best results after the combination of all acoustic features.

Another interesting evaluation would be, to see if the model predicts one of the manually assigned labels, under the assumption that *all* human labels are correct. This would reduce the need for a single-value gold standard and would give us an insight into those cases, where the model clearly misclassifies the examples.

Overall, the findings suggest, that future research should explore, whether the combination of audio and textual data on one model improves the results even more, than the combination of three separate models. Additionally, it should be further investigated whether there is an influence of the gender of the speaker on the correctness of the predictions as we did not look at it at all.

Limitations

While working with the data of the IEMOCAP data set, we realised that some of the examples are of very low quality which can negatively affect the performance of the models. This, however, does only apply to a small part of the data set. Additionally, it is not possible to correctly asses the influence of the gender on the performance. The same applies to a possible influence of the way, the data is generated: scripted or improvised.

In some models we use SELU activation function, which is still not widely used, therefore, it is possible that there are problems that are not that well known.

In general, there are limitations based on the data set. It only contains scripted and improvised recordings, by actors, which might not be representative of naturally occurring emotions. Also, as the data set is recorded in English, any generalizations outside this language are not possible.

Ethics Statement

As we work with data that has been published before the ACL Ethics Charter was implemented, we cannot guarantee that the way the data was collected and handled meets current Ethics Standards. As far as we can tell, it is still a suitable data set for this type of research. There are however the limitations mentioned above. Also, there is no information given about the age or ethnicity of the speakers.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Nikhil Chakravartula and Vijayasaradhi Indurthi. 2019. EMOMINER at SemEval-2019 task 3: A stacked

BiLSTM architecture for contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SEMEVAL 2019),* Minneapolis, Minnesota, USA, Jun 6th – 7th 2019, pages 205–209. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SEMEVAL 2019),* Minneapolis, Minnesota, USA, Jun 6th – 7th 2019, pages 39–48. Association for Computational Linguistics.

Ranjana Dangol, Abeer Alsadoon, PWC Prasad, Indra Seher, and Omar Hisham Alsadoon. 2020. Speech Emotion Recognition Using Convolutional Neural Network and Long-Short Term Memory. *Multimedia Tools and Applications*, 79(43):32917–32934.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia (MM 2010),* Firenze, Italy, Oct 25th – 29th 2010, pages 1459–1462.

Ngoc-Huynh Ho, Hyung-Jeong Yang, Soo-Hyung Kim, and Gueesang Lee. 2020. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, 8:61672–61686.

Eva Lieskovska, Maros Jakubec, and Roman Jarina. 2022. RNN with Improved Temporal Modeling for Speech Emotion Recognition. In *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*, pages 1–5. IEEE.

Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. 2021. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10(10):1163.

Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. 2019. Dying relu and initialization: Theory and numerical examples. *arXiv preprint arXiv:1903.06733*.

Shuiyang Mao, Dehua Tao, Guangyan Zhang, P. C. Ching, and Tan Lee. 2019. Revisiting Hidden Markov Models for Speech Emotion Recognition. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019),* Brighton, UK, May 12th – 17th 2019, pages 6715–6719.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer.

Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017),* New Orleans, LA, USA, Mar 5th – 9th 2017, pages 2227–2231. IEEE.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. CLaC lab at SemEval-2019 task 3: Contextual emotion detection using a combination of neural networks and SVM. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SEMEVAL 2019),* Minneapolis, Minnesota, USA, Jun 6th – 7th 2019, pages 153–158.

Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. 2003. Speech emotion recognition using hidden Markov models. *Speech communication*, 41(4):603–623.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, and Kenton Lee. 2018. Luke" Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1.

Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, Keelan Evanini, et al. 2016. The Interspeech 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), Vols 1-5,* San Francisco, USA, Sep 8th – 12th 2016, pages 2001–2005.

Thapanee Seehapoch and Sartra Wongthanavasu. 2013. Speech emotion recognition using support vector machines. In *Proceedings of the 5th international conference on Knowledge and smart technology (KST 2013),* Chonburi Province, Thailand, Jan 31th – Feb 1st 2013, pages 86–91. IEEE.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li. 2015. Speech Emotion Recognition Using Fourier Parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75.

Carl E. Williams and Kenneth N. Stevens. 1981. Vocal correlates of emotional states. *Speech evaluation in psychiatry*, pages 221–240.

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. In *Proceedings of the 2019 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP 2019), Brighton, UK, May 12th – 17th 2019, pages 2822–2826. IEEE.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT 2018)*, Athens, Greece, Dec 18th – 21th 2018, pages 112–118. IEEE.

A Computational specifications

Library	Version	Library	Version
Python	3.8.10	NLTK	3.6.7
Tensorflow	2.7.0	Seaborn	0.11.2
Keras	2.7.0	Scikit-learn	1.0.2
Tensorflow Addons	0.15.0	Matplotlib	3.5.1
Numpy	1.22.0	Librosa	0.8.1
Pandas	1.3.5		

Table 3: Libraries and their versions used in the experiments

B Method for Transcription

The model (Figure 4) starts with a vectorization layer, which is not part of the *tensorflow* or *keras* library, though it is described on their explanation page for word embeddings.[15]

C Data set

The following figures show the gender distribution of the data (Figure 5), the distribution between improvised and scripted conversations (Figure 6) and the distribution of emotion classes (Figure 7).

D Methods for audio data

The first 25 SVMs use the minimum, maximum, average and standard deviation of the listed feature (Table 6). Mirsamadi et al. (2017) trained their SVMs on these statistical features, however, they added range, skewness and kurtosis as well, which we leave out in our experiments. The classifier trained on all_frequ_ban_amp is trained on maximum, minimum, average and standard derivation of the features F1 Frequency to F3 Amplitude Log Rel F0. The classifiers all_max, all_min, all_avg and all_std are respectively trained on the maximum, minimum, average and standard deviation of the first 25 features. Classifier 31 is trained on all available features attained by using *openSMILE*.

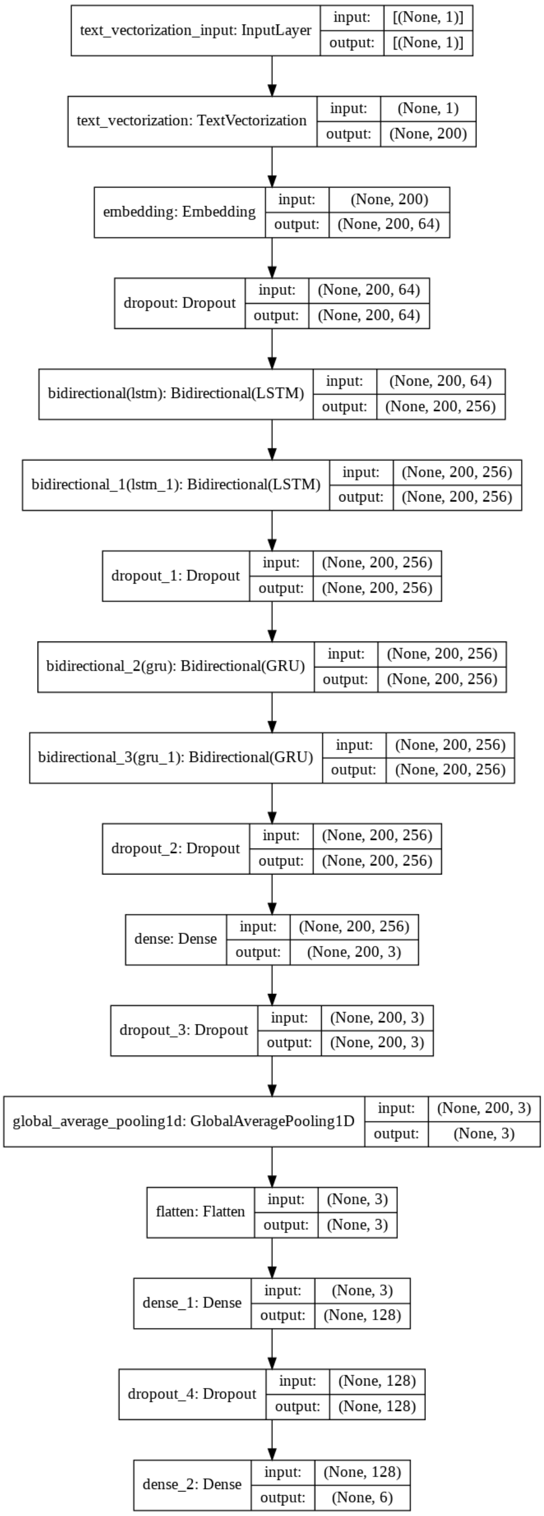


Figure 4: Architecture of the Text NN model

[15] https://www.tensorflow.org/tutorials/text/word_embeddings

Model specifics	Model name
Input layer, 3 dense layers with ReLU activation, output layer with softmax activation function	Audio NN 1 OS
Input layer, 3 dense layers with SELU activation, output layer with softmax activation function	Audio NN 2 OS
Input layer, 3 dense layers with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 3 OS
Input layer, 3 dense layers with SELU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 4 OS
2 LSTM layers, dense layer with ReLU activation, Dropout (0.2), output layer with softmax activation function	Audio NN 5 OS
2 BiLSTM layers, 2 BiGRU layers, dense layer with ReLU activation, Dropout (0.3),output layer with softmax activation function	Audio NN 6 OS
BiLSTM layer, BiGRU layer, dense layer with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 7 OS
2 BiLSTM layers, dense layer with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 8 OS
2 BiGRU layers, dense layer with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 9 OS

Table 4: Different Neural Networks trained on the features extracted via *openSMILE*. The Model name listed is the name, we use, to refer to the specific model.

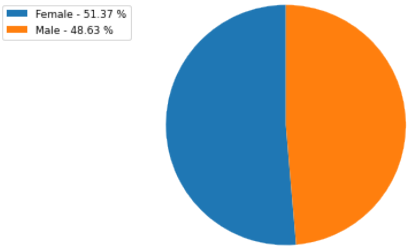


Figure 5: This Figure shows the spreading of female and male speakers on the data

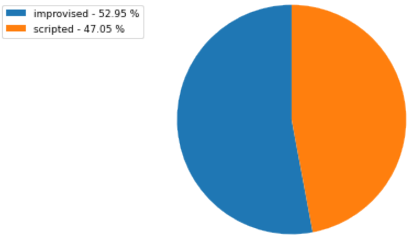


Figure 6: This Figure shows the spreading of improvised and scripted data

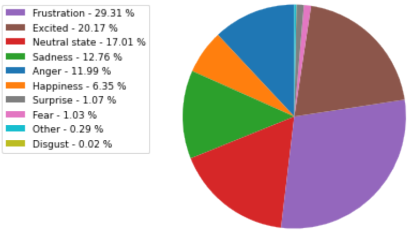


Figure 7: This Figure shows the spreading of the different classes on the data set

E Further details on results

Original Data

Class "Excited" achieves the highest precision (0.61). This aligns with the results of the combination of the Text NN with SVM SD (0.60) and the Text NN with SVM 100 (0.59). On all single models, class "Excited" reaches also the highest and second highest precision (Text NN 0.55, SVM SD 0.44 and SVM 31 0.45).

The second highest precision achieves class "Anger" (0.59), followed by classes "Frustration" and "Sadness" (both 0.52). Class "Frustration" gains the highest recall (0.72), followed by class "Excited" (0.62) and "Sadness" (0.59).

43.3% of all correctly predicted examples belong to class "Frustration", followed by class "Excited", which make up 23.8% of all correct predictions.

Data set variation 1

On the first combination, class "Happiness/Excited" reaches the highest precision of 0.85, followed by class "Anger/Frustration" (0.73) and "Sadness" (0.72). Class "Anger/Frustration" reaches the highest recall of 0.91, followed by class "Happiness/Excited" (0.65) and class "Sadness" (0.60). This leads to the highest F1 of 0.81 on class "Anger/Frustration", 0.74 on class "Happiness/Excited" and 0.65 on class "Sadness". Class "Surprise" neither has any correct predictions, nor are any examples predicted on this class at all.

Data set variation 2

Class "Anger/Frustration/Sadness" achieves the highest precision of 0.73, both other classes reach a slightly lower precision of 0.71. More differences are visible in the results of recall and F1. The highest recall (0.93) and therefore also the highest F1 (0.82) achieves class "Anger/Frustration/Sadness". Class "Happiness/Excited" reaches a recall of 0.64 and F1 of 0.67. The results on class "Neutral state, however, are even lower, with a recall of 0.20 and F1 of 0.32.

F Further combination methods

Additionally to the aforementioned combination, we combine the models based on the highest precision, recall and F1 on each class. This means, the model with the highest precision on class one is combined with the model with the highest precision on class two and so on. The same goes for recall and F1. The weighing, however, differs from the the first two approaches. In order to put more weight on the model for classes, for which it specifically gains better results, the model makes up 50% of the final prediction on this class. The other classes share the remaining 50% evenly.

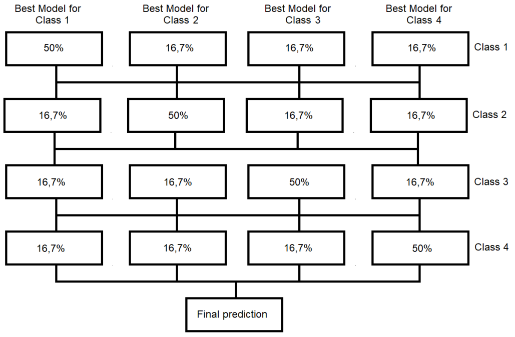


Figure 8: This Figure shows the way, the best models for each class are combined. For the class, the model performs best, it makes up 50%. The other remaining 50% percent are evenly separated between the other classes.

This means, on a data set variation with for classes (which is the case in data set variation 6 and 8), four models are combined, which can be seen in Figure 8. The best model for class one is weighing 50% in the final prediction of class one. The three models with the best performance for the other classes make up the remaining 50%. This happens for every class and on precision, recall and F1. In order to avoid to include models, which have most or even all examples of the whole data set classified on one class (recall near 1.0) or a model barely predicting a class, but if, then they are correct (precision near 1.00), we set a threshold in place. If the precision, recall of F1 is lower than 0.2, then the highest other measurement has to be lower than 0.90. This means, if, for example, the precision of a class reaches 1.00, then recall and F1 have to be at least 0.20. If not, then the model with the next lower precision is used, as long as it does not oppose to the aforementioned criteria with regard to the relation between precision, recall and F1.

Feature	Model name
Loudness	SVM 1
Alpha Ratio	SVM 2
Hammarberg Index	SVM 3
Slope 0-500	SVM 4
Slope 500-1500	SVM 5
Spectral Flux	SVM 6
MFCC 1	SVM 7
MFCC 2	SVM 8
MFCC 3	SVM 9
MFCC 4	SVM 10
F0 Semitone From 27.5Hz	SVM 11
Jitter Local	SVM 12
Shimmer Local dB	SVM 13
HNRdBACF	SVM 14
Log Rel F0-H1-H2	SVM 15
Log Rel F0-H1-A3	SVM 16
F1 Frequency	SVM 17
F1 Bandwith	SVM 18
F1 Amplitude Log Rel F0	SVM 19
F2 Frequency	SVM 20
F2 Bandwith	SVM 21
F2 Amplitude Log Rel F0	SVM 22
F3 Frequency	SVM 23
F3 Bandwith	SVM 24
F3 Amplitude Log Rel F0	SVM 25
all_frequ_ban_amp	SVM 26
all_max	SVM 27
all_min	SVM 28
all_avg	SVM 29
all_std	SVM SD
all_features	SVM 100

Table 6: Different classifiers trained on the features extracted via *openSMILE*

Model specifics	Model name
Flatten layer (input layer), 3 dense layers with ReLU activation, output layer with softmax activation function	Audio NN 1
Flatten layer (input layer), 3 dense layers with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 2
Flatten layer (input layer), 3 dense layers with SELU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 3
2 Conv2D layer (input layer), 3 MaxPooling2D layers with ReLU activation, 3 BatchNormalization layer, 1 dense layer with SELU activation, Flatten layer, Dropout (0.3), output layer with softmax activation function	Audio NN 4
2 LSTM layers, dense layer with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 5
2 BiLSTM layers, 2 BiGRU layers, dense layer with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 6
BiLSTM layer, BiGRU layer, dense layer with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 7
2 BiLSTM layers, dense layer with ReLU activation, Dropout (0.3),output layer with softmax activation function	Audio NN 8
2 BiGRU layers, dense layer with ReLU activation, Dropout (0.3), output layer with softmax activation function	Audio NN 9

Table 5: Different Neural Networks trained on the features extracted via *librosa*. The Model name listed is the name, we use, to refer to the specific model.