Technische Hochschule Ingolstadt

Fakultät Informatik
Bachelorstudiengang Künstliche Intelligenz

Evaluating Generalization Capabilities of a Speech Emotion Recognition System on Acted and Semi-Natural Datasets

Seminararbeit

Philip Weinmann

Betreuung:	Prof. Munir Georges, Mariano Frohnmaier
Datum:	July 1, 2022

Abstract

Speech emotion datasets can be divided into three core categories: acted, semi-natural/semi-simulated and natural databases. Most datasets are very small and biased resulting in various challenges for Speech Emotion Recognition (SER) research. Current SER research mostly uses acted and semi-natural datasets to train and evaluate their models due to the lack in natural emotional speech datasets. Acted emotional speech datasets are comparably easy to create and can be designed in a very machine-learning friendly fashion (standardized sample duration, predefined target emotions). Yet the problem of using acted emotional speech is that it is perceived exaggerated and unnatural by humans. Nevertheless many researches are based on this acted emotional speech data. In this work the generalization capabilities of a model that was trained on an acted speech dataset is evaluated by testing the model on a semi-simulated dataset. The results show that the model trained on acted emotional speech is not capable of generalizing to semi-natural speech. In another model the experiment was mirrored by evaluating a model trained on semi-natural speech, on an acted speech data. The second experiment delivered results, that are worthy of furhter investigation and discussion.

Contents

1 Introduction **1**

2 Speech Emotion Data **2**

3 Experiments **5**

3.1 Speech Features 5

3.2 Experimental setup 6

3.3 Evaluation metrics 9

4 Experiment results **10**

4.1 Current State of the Art 10

4.2 Results and interpretation 11

5 Conclusion **14**

Literaturverzeichnis **15**

A Abbreviations **19**

B Appendix **19**

1 Introduction

Humans have evolved various forms of communication like speech, facial expressions, gestures or body postures, with speech being one of the fastest and most natural ones. All these forms of human communication carry information at two levels: the message itself and the underlying emotional state. [1]

Automatic Speech Recogntion (ASR) systems try to recognize the message of the given speech by converting it into text. A popular method of evaluation is the Word Error Rate (WER) [2] that publications try to lower further and further. This research field was able to benefit a lot from the latest rising era of deep learning and achieved astonishing results that have found their way into millions of homes or pockets with products like Amazon Alexa or Apple Siri.

SER research on the other hand tries to recognize different emotional states from speech. This research area was not able to achieve comparable results despite the developments in deep learning. Some of the reasons for this are the lack of expository emotional speech data, resulting from the subjectivity in annotating data, the comparably strong effort that is needed for creating emotional speech or the challenge of combining the most fitting speech features for SER-models.

Since the ability to work on large data scales and to extract features from data without manual support are big strengths of deep learning techniques [3], the lack of emotional speech data makes it particularly hard for SER systems to benefit from deep learning developments and to create models with high generalization capabilities.

Current SER frameworks that deliver state of the art results are the shallow Convolutional neural network (CNN) model mentioned in [4] that achieves an accuracy of about 82% on the Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess) dataset and the HuBERT [5] based model by IBM [6] that achieves similar results on the Interactive emotional dyadic motion capture database (Iemocap) dataset.

SER datasets are categorized into natural, semi-natural and simulated, with most of them being in the simulated category[7]. The fundamental difference between these categories is the process in which the emotion within the speech was created. Two of the most common datasets that are being used very frequently for SER training and benchmarks are the Ravdess [8] and the Iemocap [9], with Ravdess representing the simulated and Iemocap representing the semi-simulated category.

The goal of this work is to test how well a transformer [10] and CNN based framework

[11], that is evaluated on the acted and the semi-simulated domain beforehand, performs on Iemocap after being trained on Ravdess. This allows to make assumptions about the quality of the simulations of the emotional speech and the generalization capabilities of models that have been trained on simulated emotional speech datasets.

2 Speech Emotion Data

Speech emotion datasets can be divided into three categories: simulated, semi-simulated and natural. The fundamental difference between these categories is the process in which the emotion in the speech is created.

Natural datasets are usually extracted from video and audio clips like YouTube-videos, podcasts or TV-shows. It can definitely be questioned if emotional speech in YouTube-videos, podcasts or TV-shows represents natural emotional speech and should be considered 'natural' emotion but this work and its experiments will be focused on acted and semi-natural datasets. Nevertheless in theory while creating SER architectures, one would intuitively choose large-scale natural datasets to train these models on, because natural data represents real emotional speech best and should thus give large generalization capabilities. Yet there are some constraints, that make natural emotional speech datasets rare. As stated in [7], the use of natural speech for SER models is complicated due to the existence of concurrent emotions together, the contentiousness of emotion and the presence of noise. This results in a high effort that needs to be done manually in pre-pocessing pipelines for SER models. Another problem is that emotion is rare in natural speech. In most everyday situations the neutral emotion will be the dominant one, making datasets very imbalanced. [12]

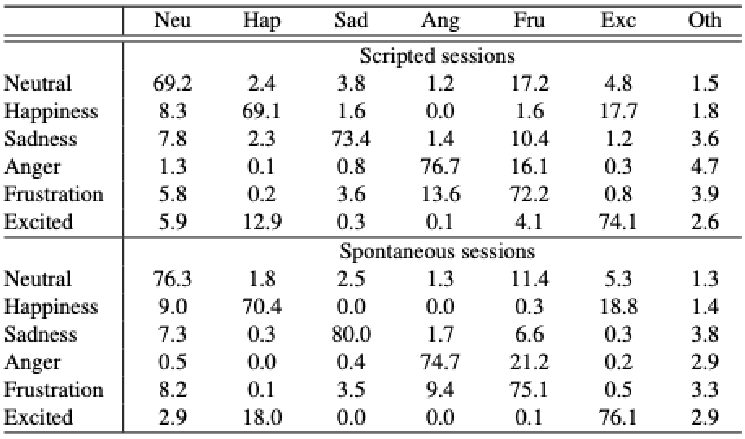
Burkhardt et al. resumed the using of natural speech emotion data as follows: 'As clear emotional expression is not only rare in everyday situations but also the recording of people experiencing full-blown emotions is ethically problematic, it is almost impossible to use natural data if basic emotions are the subject of investigation.' [13] As a result, natural emotional speech data as well as other emotional speech data is very sparse and thus insufficient in diversity. One promising project that should be mentioned is the MSP-Podcast corpus[14] . The project aims to build the largest natural speech emotion dataset and could be promising for the creation of generalising SER models.

Simulated emotional speech is created by trained speakers who read text with different emotions that are previously defined.[7] One of the most famous representatives of this category is Ravdess. The message in the text sample that is read by the actors intentionally does not contain any emotional hints (e.g. text sample from Ravdess: 'dogs are sitting by the door.'), so the actors focus on their voice volume or prosody to display the desired emotion. The main advantages of this approach are that it is very intuitive, relatively easy to carry out and that the labels for the speech samples are clear upfront and accordingly to this do not have to be created afterwards in a time consuming process.

On the other hand the simulation of emotional speech has two main drawbacks: First, as the authors of Iemocap stated, there is no guarantee for the recorded utterance to match the target emotion. The reason for this is that emotional speech is expressed different by different speakers, due to speaker variation and additionally emotional speech is perceived different by different listeners. Many papers show the importance of speaker variation, for example a study of Sethu et al. has shown that 'the accuracy's of emotion [classification] is affected to a larger extend by differences between speakers than they are by difference between phonemes' [15]. One of the latest SER frameworks, developed by IBM [6], was explicitly set up in a fashion that deals with the speaker variation problem and achieved state of the art (SOTA) results on Iemocap. This proves how big of an influence speaker variation has on the quality of SER. A intuitive solution to deal with both speaker variation and different emotional perceptions is to use large scale datasets that are big enough to cover most of the variance in human ways of expressing emotions and to annotate them by a wide range of people from different sociocultural backgrounds [16]. Second, simulated emotional speech causes a more extreme perception of the displayed emotion, leading to unnaturalness. This is quite obvious when listening to the samples of Ravdess and was proven in a study by J. Wilting et al. [17] in which he found out that acted emotion is perceived more extreme and thus partly unnatural.

Iemocap is the most widely used semi-simulated dataset. The emotional speech is created by ten actors in two ways: scripted sessions and spontaneous sessions. In the scripted sessions actors were asked to memorize a script with a length of about ten minutes. These scripts were chosen by a theater professional given that the plays should contain target emotions and that each play consists of one female and one male role. For the spontaneous sessions the actors were given hypothetical scenarios which were designed to contain the target emotions. These scenarios were chosen by the guidelines of Scherer et al. [18], who polled participants who were asked to remember scenarios in which they felt certain emotions. Examples for these scenarios are: the loss of a friend or separation.

In these semi-simulated creation procedures it is necessary to annotate the recorded utterances afterwards, because actors are not given clear target emotions but the freedom to play their scripts or scenarios based on their own subjective interpretation. As mentioned above, the choice and the number of annotators have an influence on the performance and the generalization capabilities of models that are trained on the dataset and thus should be chosen carefully. At best, a large sample of people representing many different cultures, backgrounds, ages, countries, etc. should be annotating the data. For Iemocap six annotators evaluated the scripted and spontaneous sessions. The final ground truth per utterance was then defined by their majority vote.



	Neu	Hap	Sad	Ang	Fru	Exc	Oth
	Scripted sessions						
Neutral	69.2	2.4	3.8	1.2	17.2	4.8	1.5
Happiness	8.3	69.1	1.6	0.0	1.6	17.7	1.8
Sadness	7.8	2.3	73.4	1.4	10.4	1.2	3.6
Anger	1.3	0.1	0.8	76.7	16.1	0.3	4.7
Frustration	5.8	0.2	3.6	13.6	72.2	0.8	3.9
Excited	5.9	12.9	0.3	0.1	4.1	74.1	2.6
	Spontaneous sessions						
Neutral	76.3	1.8	2.5	1.3	11.4	5.3	1.3
Happiness	9.0	70.4	0.0	0.0	0.3	18.8	1.4
Sadness	7.3	0.3	80.0	1.7	6.6	0.3	3.8
Anger	0.5	0.0	0.4	74.7	21.2	0.2	2.9
Frustration	8.2	0.1	3.5	9.4	75.1	0.5	3.3
Excited	2.9	18.0	0.0	0.0	0.1	76.1	2.9

Figure 1: confusion matrices from individual subjective perceptions compared to ground truths. Ground truth defined by majority vote. [9]

Conclusive, when choosing a dataset a trade off between the naturalness of the speech emotion, the amount of effort needed to preprocess the data and the size of the dataset has to be made. As mentioned above, natural data is being ethically and legally problematic while simulated data is unnatural, leading to a lack of data in general and hence overfitting problems in most SER approaches [19].

3 Experiments

The goal of this work was to find out how well a SER model that was trained on the acted Ravdess would perform on the semi-natural domain of the Iemocap, to help answering the following questions:

- are SER models trained on simulated data capable of generalising on more natural domains?
- is it useful to build SER architectures that are explicitly fit to simulated datasets?

The model I used for this experiment was proposed by Ilya Zenkov in 2020. [20] With an unweighted accuracy (UA) of 80.44% on Ravdess this architecture still competes with SOTA results. The architecture consists of two different CNN's and a transformer with multi-head attention (see equation 1 and 2).

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where $\sqrt{d_k}$ is the dimension of the key vector k and query vector q

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1, ..., head_h)W^O \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2}$$

The models respectively extract feature maps from the input feature in parallel, as shown in figure 2. The resulting feature maps are concatenated and flattened into a dense, latent vector that is fed into a final linear layer. The class probabilities are calculated with the softmax function:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K \tag{3}$$

In the following sections I will describe the choice of speech features, the experimental setup and the evaluation metrics.

3.1 Speech Features

There are many studies suggesting different combinations of speech features. With the latest rise of the deep learning era most of the recent works in SER only use one to three

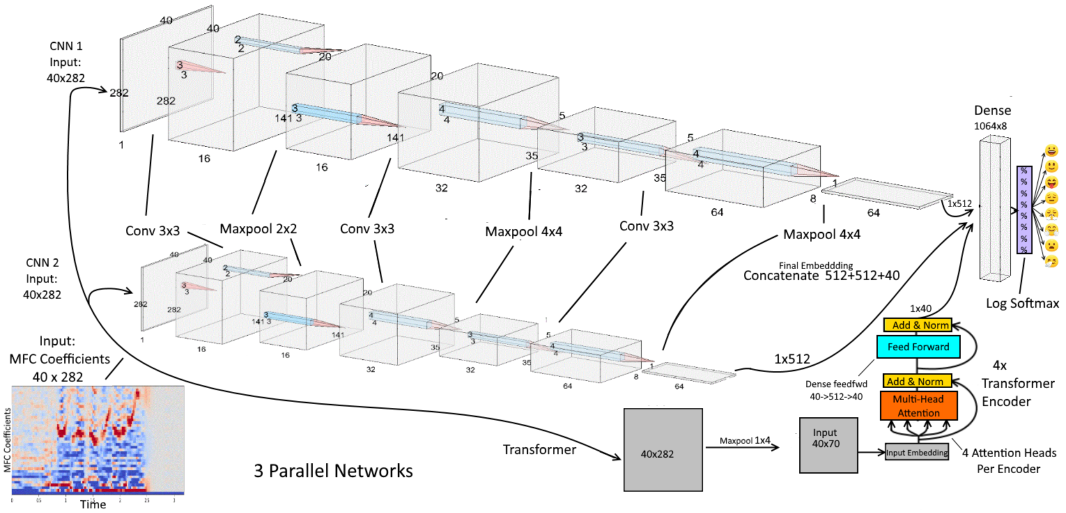


Figure 2: Zenkov's 'parallel is all you need' architecture [11]

speech features [7]. In traditional machine learning methods it was necessary to handcraft representative speech features since there was no alternative. Most traditional approaches used a combination of many different spectral and temporal speech features. This required expert knowledge and made the speech features strongly domain dependant. In deep learning on the other hand one can make use of automatically extracted features by the network. Therefore the features can become domain independent and, based on the architecture, translation invariant [21].Table 1 shows a small excerpt of more recent SER - research titles with the features they used.

It is noticeable that most researches focus on very few features due to the reasons stated previously. Zenkov additionally evaluated his framework on different feature combinations but found that the best results are achieved when only using mel-frequency cepstral coefficients (MFCC). The feature used for the upcoming models is therefore also only MFCC. With the librosa library [26] 40 MFCC coefficients were extracted per speech sample.

3.2 Experimental setup

The experiment is to train and validate a model on Ravdess and evaluate it on Iemocap. In order to put these results into perspective and to be able to interpret them, I considered it necessary to create a total of 4 models with different training-, validation- and test-dataset splits as shown in table 2:

Research Title	Features
Improved end-to-end Speech Emotion Recognition using self-attention mechanism and multitask learning, Li et al. [22]	Mel-Scale-Spectogram
Negative Emotion Recognition using Deep Learning for Thai Language, Mekruksavanich et al. [23]	MFCC
Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition, Eskimez et al. [24]	Log-Mel-Spectogram
Speech Emotion Recognition using deep 1D and 2D CNN LSTM networks, Zhao et al. [25]	PCM, Log-Mel-Spectogram
Parallel is all u need, Zenkov [11]	MFCC

Table 1: Recent models and their speech feature choice

Model 1 and 2 (rav-only and iemo-only) were created to evaluate the used network architecture itself. By doing this, the results can be compared to the respective SOTA framework that are presented in section 4.1. Model 3 and its results build the foundation of this experiment. The results are crucial for the considerations made about the generalization capabilities of SER model into the semi-natural domain, after training on an acted dataset. The last model adds more room for interpretation since it mirrors the experiment done by model 3. It allows assumptions about the generalization capabilities into the acted domain, after training on semi-natural data.

In every model the cross entropy $L(y_{o,c}, p_{o,c}) = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$ is used as a loss function. Tripathi et al. [27] were able to achieve better results on their SER framework by using a focal loss instead of the typical cross entropy, but for the experiments in this work, no additional improvements were made by switching from a cross entropy to a focal loss.

k	Model	Train-set	Validation-set	Test-set
1	rav-only	Ravdess 80%	Ravdess 10%	Ravdess 10%
2	iemo-only	Iemocap 80%	Iemocap 10%	Iemocap 10%
3	iemo-test	Ravdess 90%	Ravdess 10%	Iemocap 100%
4	rav-test	Iemocap 90%	Iemocap 10%	Ravdess 100%

Table 2: dataset splits per model

To keep the conditions as similar as possible for every experiment, the same hyperparameters (table 3) were used for every model. Due to the long training duration, no extensive hyperparameter optimization was done. Only different learning rates (0.0001, 0.001, 0.01) and batch sizes (32, 64, 128) were tested, with the values shown in table 3 achieving the highest results.

batch-size	$b = 32$
learning-rate	$\alpha = 0.01$
momentum coefficient	$\lambda = 0.8$
weight-decay	$\gamma =$ 1e-3
dropout-rate CNN	$p_{cnn}(n) = 0.3$
dropout-rate transformer	$p_{attention}(n) = 0.4$
CNN layers	3 per CNN
self-attention heads	4

Table 3: Chosen hyperparameters for the experiments

Every model was trained to classify 4 emotions (neutral, happy, sad, angry). Therefore the number of emotions had to be reduced from 8 to 4 on Ravdess and from 11 to 4 on Iemocap. The reason for this reduction is, that Ravdess contains emotions that Iemocap doesn't contain and vice versa. Additionally, due to Iemocap being a semi-natural database, its class counts are not balanced and some of them are very sparse, as can be seen in table 4.

The models were trained for 300 epochs each. They all started overfitting in the first 50

Emotion	Count Ravdess	Count Iemocap
surprised	192	107
neutral	96	1708
calm	192	NaN
happy	192	595
sad	192	1094
angry	192	1103
fearful	192	40
disgust	192	2
other	NaN	3
frustrated	NaN	1849
excited	NaN	1041

Table 4: Emotion Counts

epochs and thus were saved as checkpoints after every epoch. Finally, the model with lowest validation error before overfitting on the training data, was chosen for the final evaluations on the test-set.

3.3 Evaluation metrics

The UA (see equation 4) is a very popular benchmark performance indicator in SER, used in many publications. Regardless of the number of samples per class, the influence on the accuracy score is the same for all classes. This leads to give more weight to low sample classes and reduce the weight of high sample classes. For this reasons, a normalized confusion matrix and the UA are calculated for every model.

$$\frac{1}{C} \cdot \sum_{c=1}^{C} \frac{\text{correct predictions}_c}{\text{all predictions}_c} \quad \text{for Classes 1, ..., C} \tag{4}$$

To present the relation between precision and recall in the confusion matrices, the F1-score 5 is also calculated to evaluate the experiments. It can be interpreted as an harmonic mean of precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

4 Experiment results

In this section the results of the experiments are presented. To be able to interpret the results better, I will first present two research titles that obtained SOTA results on Ravdess and Iemocap respectively.

4.1 Current State of the Art

With 'Speaker Normalization for Self-Supervised Speech Emotion Recognition' [6] Gat et al. published a new SOTA framework on the semi-natural Iemocap. As mentioned in chapter 3, speaker variation is a problem for SER because when dealing with deep neural networks and small datasets, networks tend to find shortcuts by identifying the speaker. Therefore Gat et al. chose an approach in which the framework is prevented to do this. Their model consists of a HuBERT Large [5] backbone, that they call 'upstream model' and that generates rich semantic feature representations from the speech samples. The backbone is connected to two 'downstream models': a speaker identification (SID)-model and a SER-model. By training the SID-model on a large scale dataset and the accompanying learning of speaker characteristics, they are able to invert the speaker information for the emotion recognition task by simply negating the gradients from the SID model in the process of back-propagation. The approach achieved SOTA results of 81% UA on Iemocap while using the balanced classes: happy, neutral, sad and angry.

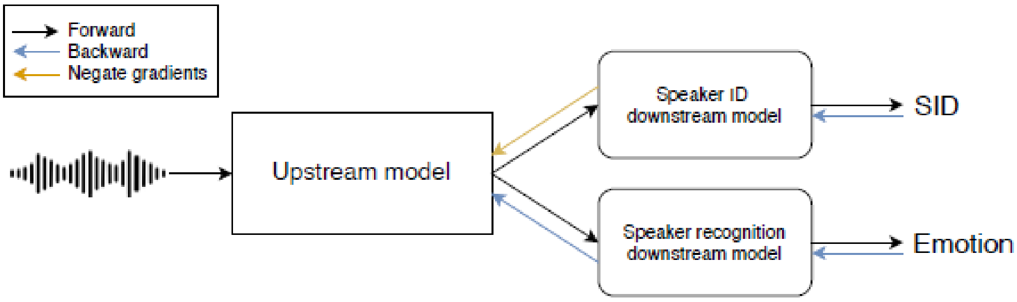

Forward
Backward
Negate gradients
Upstream model
Speaker ID downstream model
Speaker recognition downstream model
SID
Emotion

Figure 3: Architecture of the speaker normalization framework [6]

In their research 'Shallow over Deep Neural Networks: A Empirical Analysis for Human Emotion Classification Using Audio Data' [4], Kanani et al. developed a SER framework on Ravdess. In comparison to Gat et al. they were using all 8 emotion-classes in the dataset, since Ravdess is an acted dataset and therefore already created fairly balanced. In a first step they convert the raw audio waves into a Mel Spectogram image of size 224x224. By using only the Mel Spectogram they changed the SER task into a computer vision problem. The tested and compared different popular image classification models, like VGG-16 [28] or Inception V3 [29] on the emotion recognition task. Ultimately they created a new architecture that they call 'CNN-X'. By 3 convolutional blocks, consisting of an 3x3 convolution layer and an 2x2 average pooling layer, they reduced the spatial resolution of the input spectograms to 28x28 while stacking 32 feature channels. The feature map is flattened and fed into 2 fully connected layers to classify the emotions (see figure 4). With their approach they achieved results of 82.99 % weighted accuracy. To my best knowledge, this is SOTA on the 8 class classification task on Ravdess. At the time of writing this work, the website 'paperswithcode.com' [30] also ranks them as SOTA.

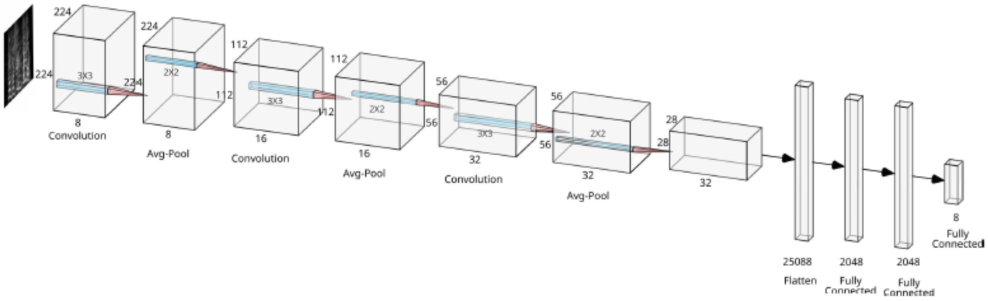


Figure 4: Shallow CNN architecture [4]

4.2 Results and interpretation

Through this section, the main outcomes obtained from the experiments will be presented. In table 2 the training, validation and test splits per model as well as the model names used in this section are defined. Table 5 summarizes the results of the 4 different SER models.

Model	UA	F1-Score (weighted)
rav-only	98.5 %	98.5 %
iemo-only	68.5 %	72.6 %
iemo-test	33.75 %	33.9 %
rav-test	31.0 %	31.0 %

Table 5: dataset splits per model

Comparing the rav-only and the iemo-only model shows the differences in difficulty of SER on a semi-natural dataset versus SER on an acted dataset. The confusion matrices for these two models can be found at the link to the github (see appendix). The UA dropped by 30 % from switching the dataset category. When recreating the experiment of Ilia Zenkov I achieved an UA of 69 % on Ravdess while classifying all 8 emotions in the dataset, which is still more accurate than the results of the 4 class iemo-only model. This confirms the statements about semi-natural and acted datasets, made in section 2: The semi-natural database is more complex, less standardized and more difficult to preprocess in comparison to the acted dataset. This can also be concluded by the fact, that the two SOTA models, presented in section 4.1, achieve almost similar accuracy, with the model trained on Iemocap only being able to classify 4 emotions while the model trained on Ravdess is able to classify 8 emotions. So even with best pre-processing and expert knowledge there will be a trade-off between the accuracy and the naturalness of the emotional speech.

The results of the iemo-test model show that even though a SER framework delivers astonishing results when performing on an acted dataset, the same model is not able to generalise to a semi-natural domain. With an UA of 33.75 %, the model is barely better than a random classifier. Figure 5 shows the confusion matrix of the model on Iemocap. It is notable that the model strongly tends to the classes 'neutral' and 'angry' and that the 'sad' emotion got recognized correctly in only 2 % of the cases. This model and its results illustrate the strong differences between the two datasets and the difference between acted emotion to semi-natural emotion.

The last model that was trained, the rav-test model, mirrored the actual experiment. By

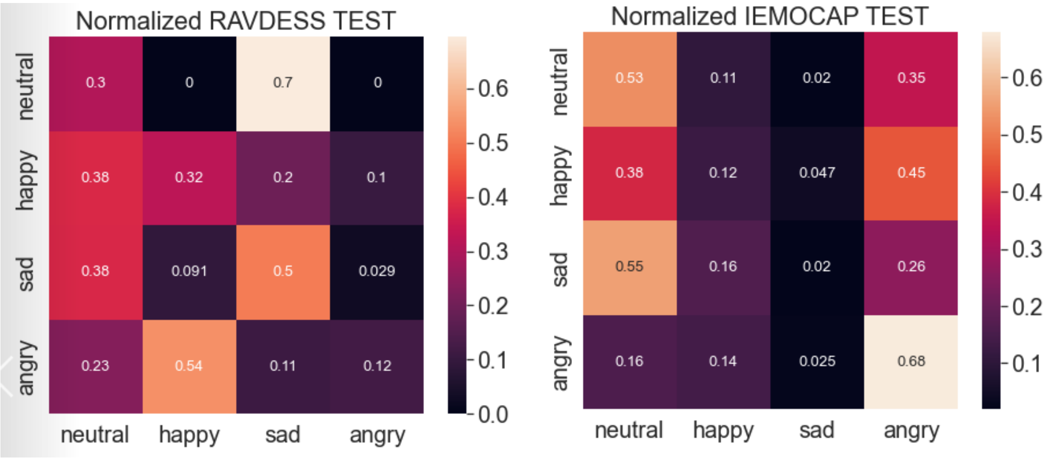


Figure 5: Confusion matrices for the rav-test and the iemo-test models

training in semi-natural data and testing on the acted data, the goal was to show, that semi-natural data are able to generalise to an acted domain. With the UA for this model achieving 31 %, this could not be proven in this specific experiment.
The results can be interpreted in different ways:

- The differences between the two datasets are to big and they can thus be seen as completely different domains.
- The models were not able to learn actual emotional representations but maybe took other shortcuts like identifying the speakers in the datasets and therefore have problems with generalising to other datasets.
- Models trained on semi-natural datasets are better in generalising towards the acted emotional speech domain.

The last interpretation can be explained the following: The comparison between the rav-only and iemo-only models showed that acted speech samples are recognized significantly better than semi-natural speech samples. Nevertheless, the results when testing the generalization capabilities (iemo-test and rav-test) were almost identical. When considering the UA that a model achieved, when tested on the domain it was trained on, into its generalization capabilities, for example by taking the fraction of the UA's, the following figures result:

$$gc_{ravdess-trained-models} = \frac{\mathrm{UA}_{iemo-test}}{\mathrm{UA}_{rav-only}} = \frac{33.75\%}{98.5\%} = 34.3\%$$

$$gc_{iemocap-trained-models} = \frac{\mathrm{UA}_{rav-test}}{\mathrm{UA}_{iemo-only}} = \frac{31\%}{68.5\%} = 45.3\%$$

In this case the Iemocap trained model scores higher than the Ravdess trained model even though the architecture itself and the choice of features was originally built and optimized for the Ravdess dataset. Still, due to the low UA of the rav-test and the iemo-test model, their results are not overly expressive and more investigation must be done before making general assumptions at this point.

5 Conclusion

In this work I presented an experiment to investigate the generalization capabilities of a SER framework, that is trained on acted emotional speech, to a semi-natural emotional speech domain. The main goal was to prove that acted emotional speech models are not able to generalize, to prevent extensive researches and benchmarks on acted emotional datasets, due to their lack of usability for real world SER. In the approach I first evaluated the network architecture itself by training and testing exclusively on acted and semi-natural speech data. In a next step, a model that was trained and validated on an acted dataset, was tested on semi-natural emotional speech data. The model achieved an UA of 33.75 % when classifying 4 emotions and therefore is barely better than a random classifier, proving the initial goal of this work. I therefore recommend not to use acted speech data when training SER models and to invest effort into pre-processing and fine-tuning more natural emotional speech data.

When mirroring the experiment and training on semi-natural data and testing on acted data, the results were almost equal. The fact, that the model that was trained and tested on Ravdess achieved significantly better results than the model trained and tested on Iemocap, was lastly taken into account to make a final comparison between the generalization capabilities. Even though the semi-natural model has an higher relational score, conclusions about the generalization capabilities of semi-natural based models need further investigations and tests.

References

[1] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, “Analysis of emotional speech—a review,” *Toward robotic socially believable behaving systems-volume i*, pp. 205–238, 2016.

[2] Y. Park, S. Patwardhan, K. Visweswariah, and S. Gates, “An empirical analysis of word error rate and keyword error rate,” 09 2008, pp. 2070–2073.

[3] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[4] C. S. Kanani, K. S. Gill, S. Behera, A. Choubey, R. K. Gupta, and R. Misra, “Shallow over deep neural networks: A empirical analysis for human emotion classification using audio data,” in *International Conference on Internet of Things and Connected Technologies*. Springer, 2020, pp. 134–146.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[6] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, “Speaker normalization for self-supervised speech emotion recognition,” *arXiv preprint arXiv:2202.01252*, 2022.

[7] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep learning techniques for speech emotion recognition, from databases to models,” *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[8] S. R. Livingstone and F. 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*, vol. 13, no. 5, p. e0196391, 2018.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] I. Zenkov, "transformer-cnn-emotion-recognition," https://github.com/IliaZenkov/transformer-cnn-emotion-recognition, 2020.

[12] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[14] https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html, timestamp: 19.06.2022-14:53.

[15] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and speaker variations in automatic emotion classification," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[16] H.-C. Chou and C.-C. Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5886–5890.

[17] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech." in *Interspeech*, vol. 2006, 2006, p. 9th.

[18] H. F. Wallbott *et al.*, *Experiencing emotion: A cross-cultural study.* Cambridge University Press, 1986.

[19] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "Lssed: a large-scale dataset and benchmark for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 641–645.

[20] B. T. Atmaja, K. Shirai, and M. Akagi, “Speech emotion recognition using speech feature and word embedding,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 519–523.

[21] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, “Deep learning approaches for speech emotion recognition: State of the art and research challenges,” *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23 745–23 812, 2021.

[22] Y. Li, T. Zhao, and T. Kawahara, “Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.” in *Interspeech*, 2019, pp. 2803–2807.

[23] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, “Negative emotion recognition using deep learning for thai language,” in *2020 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON)*. IEEE, 2020, pp. 71–74.

[24] S. E. Eskimez, Z. Duan, and W. Heinzelman, “Unsupervised learning approach to feature analysis for automatic speech emotion recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5099–5103.

[25] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1d & 2d cnn lstm networks,” *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[26] https://librosa.org/doc/latest/index.html, timestamp: 24.06.2022-09:47.

[27] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, “Focal loss based residual convolutional neural network for speech emotion recognition,” *arXiv preprint arXiv:1906.05682*, 2019.

[28] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[30] https://paperswithcode.com/sota/speech-emotion-recognition-on-ravdess, timestamp: 29.06.2022-10:55.

A Abbreviations

SER Speech Emotion Recognition

ASR Automatic Speech Recogntion

WER Word Error Rate

Ravdess Ryerson Audio-Visual Database of Emotional Speech and Song

Iemocap Interactive emotional dyadic motion capture database

CNN Convolutional neural network

SOTA state of the art

MFCC mel-frequency cepstral coefficients

UA unweighted accuracy

SID speaker identification

B Appendix

The code to the experiments as well as reports can be found at the following github-repository:
https://github.com/flippi247/Speech-Emotion-Recognition

Eidesstattliche Erklärung

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Philip Weinmann

Ingolstadt, July 1, 2022

Unterschrift