The MuSe 2022 Multimodal Sentiment Analysis Challenge: 
Humor, Emotional Reactions, and Stress

Lukas Christ
University of Augsburg
Augsburg, Germany

Shahin Amiriparian
University of Augsburg
Augsburg, Germany

Alice Baird
Hume AI
New York, USA

Panagiotis Tzirakis
Hume AI
New York, USA

Alexander Kathan
University of Augsburg
Augsburg, Germany

Niklas Müller
University of Passau
Passau, Germany

Lukas Stappen
Recoro
Munich, Germany

Eva-Maria Meßner
University of Ulm
Ulm, Germany

Andreas König
University of Passau
Passau, Germany

Alan Cowen
Hume AI
New York, USA

Erik Cambria
Nanyang Technological University
Singapore

Björn W. Schuller
Imperial College London
London, United Kingdom

ABSTRACT

The Multimodal Sentiment Analysis Challenge (MuSe) 2022 is dedicated to multimodal sentiment and emotion recognition. For this year’s challenge, we feature three datasets: (i) the Passau Spontaneous Football Coach Humor (Passau-SFCH) dataset that contains audio-visual recordings of German football coaches, labelled for the presence of humour; (ii) the HUME-REACTION dataset in which reactions of individuals to emotional stimuli have been annotated with respect to seven emotional expression intensities, and (iii) the Ulm-Trier Social Stress Test (ULM-TSST) dataset comprising of audio-visual data labelled with continuous emotion values (arousal and valence) of people in stressful dispositions. Using the introduced datasets, MuSe 2022 addresses three contemporary affective computing problems: in the Humor Detection Sub-Challenge (MuSe-HUMOR), spontaneous humour has to be recognised; in the Emotional Reactions Sub-Challenge (MuSe-REACTION), seven fine-grained ‘in-the-wild’ emotions have to be predicted; and in the Emotional Stress Sub-Challenge (MuSe-STRESS), a continuous prediction of stressed emotion values is featured. The challenge is designed to attract different research communities, encouraging a fusion of their disciplines. Mainly, MuSe 2022 targets the communities of audio-visual emotion recognition, health informatics, and symbolic sentiment analysis. This baseline paper describes the datasets as well as the feature sets extracted from them. A recurrent neural network with LSTM cells is used to set competitive baseline results on the test partitions for each sub-challenge. We report an Area Under the Curve (AUC) of .8480 for MuSe-HUMOR; .2801 mean (from 7-classes) Pearson’s Correlations Coefficient (ρ) for MuSe-REACTION, as well as .4931 Concordance Correlation Coefficient (CCC) and .4761 for valence and arousal in MuSe-STRESS, respectively.

CCS CONCEPTS

• Computing methodologies → Neural networks; Artificial intelligence; Computer vision; Natural language processing.

KEYWORDS

Multimodal Sentiment Analysis; Affective Computing; Humor Detection; Emotion Recognition; Multimodal Fusion; Challenge; Benchmark

ACM Reference Format:

Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. 2022. The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress. In Proceedings of the 3rd International Multimodal Sentiment Analysis Workshop and Challenge (MuSe’ 22), October 10, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3551876.3554817

1 INTRODUCTION

The 3rd edition of the Multimodal Sentiment Analysis (MuSe) Challenge addresses three tasks: humour detection and categorial as well as dimensional emotion recognition. Each corresponding sub-challenge utilises a different dataset. In the Humor Detection Sub-Challenge (MuSe-HUMOR), participants will detect the presence of humour in football press conference recordings. For MuSe-HUMOR, the novel Passau Spontaneous Football Coach Humor (Passau-SFCH) dataset is introduced. It features press conference recordings of 10 German Bundesliga football coaches, recorded between August 2017 and November 2017. Initially, the dataset comprises about 18 hours of video, where each of the 10
Table 1: Reported are the number (#) of unique subjects, and the duration for each sub-challenge hh:mm:ss.

| Partition    | MuSe-Humor Duration | MuSe-Reaction Duration | MuSe-Stress Duration |
|--------------|----------------------|-------------------------|----------------------|
| Train        | 3:52:44              | 3:51:04:02              | 3:25:56              |
| Development  | 3:58:12              | 444                    | 14:59:27             |
| Test         | 3:55:41              | 444                    | 14:48:21             |
| Total        | 10:59:37             | 2222                   | 74:26:19             |

The paper is structured as follows: Section 2 introduces the three sub-challenges alongside with the datasets they are based on, and outlines the challenge protocol. Then, pre-processing, provided features, their alignment, and our baseline models are described in Section 3. In Section 4, we present and discuss our baseline results before concluding the paper in Section 5.

A summary of the challenge results can be found in [1].

2 THE THREE SUB-CHALLENGES

In what follows, each sub-challenge and dataset is described in detail, as well as the participation guidelines.

2.1 The MuSe-Humor Sub-Challenge

Humour is one of the richest and most consequential elements of human behaviour and cognition [25] and thus of high relevance in the context of affective computing and human-computer interaction. As humour can be expressed both verbally and non-verbally, multimodal approaches are especially suited for detecting humour. However, while humour detection is a very active field of research in Natural Language Processing (e.g., [16, 52]), only a few multimodal datasets for humour detection exist [28, 38, 51]. Especially, to the best of our knowledge, there are no datasets for detecting humour in spontaneous, non-staged situations. With MuSe-Humor, we intend to address this research gap.

In this challenge, the PASSAI-SFCH dataset is utilised. It features video and audio recordings of press conferences of 10 German Bundesliga football coaches, during which the coaches occasionally express humour. The press conferences present natural, live, semi-staged communication of the coaches to and with journalists in the audience. All subjects are male and aged between 30 and 53 years. The dataset is split into speaker independent partitions. The training set includes the videos of 4 coaches, while the development and test partition both comprise the videos of 3 coaches.

By providing the mentioned tasks in the 2022 edition of MuSe, we aim for addressing research questions that are of interest to affective computing, machine learning and multimodal signal processing communities and encourage a fusion of their disciplines. Further, we hope that our multimodal challenge can yield new insights into the merits of each of the core modalities, as well as various multimodal fusion approaches. Participants are allowed to use the provided feature sets in the challenge packages and integrate them into their own machine learning frameworks.

1https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german
Every video was originally labelled by 9 annotators at a 2 Hz rate indicating sentiment and direction of the humour expressed, as defined by the two-dimensional humour model proposed by Martin et al. [36] in the Humor Style Questionnaire (HSQ). For the challenge, we only build upon binary humour labels, i.e., indicating if the coach’s communication is humorous or not. We obtain a binary label referring to presence or absence of humour using the following three steps. First, we only consider the humour dimension label of sentiment. Second, based on the sentiment labels, we filter out annotators displaying low agreement with other annotators. In order to account for slight lags in annotation signals, we choose to compute the target humour labels for frames of two seconds using a step size of one second. Finally, such a frame is considered as containing humour if at least 3 of the remaining annotators indicate humour within this frame. As a result, 4.9% of the training partition frames, 2.9% of the development partition frames, and 3.9% of the test partition frames are labelled as humorous. We deliberately opted for a split in which the humour label is over-represented in the training partition in order to help participants’ models with learning. The provided features are extracted at 2 Hz rates. They can easily be mapped to the 2 s segments they belong to.

For evaluation, the AUC metric is utilised, indicating how well a model can separate humorous from non-humorous frames.

2.2 The MuSe-Reaction Sub-Challenge

Computational approaches for understanding human emotional reactions are of growing interest to researchers [32, 46], with emerging applications ranging from pedagogy [15] to medicine [41]. A person’s reaction to a given stimulus can be informative about both the stimulus itself, e.g., whether educational material is interesting to a given audience, and about the person, e.g., their level of empathy [48] and well-being [57]. However, progress in developing computational approaches to understand human emotional reactions has been hampered by the limited availability of large-scale datasets of spontaneous emotional reactions. Thus, for the MuSe-Reaction sub-challenge, we introduce the Hume-Reaction (HUME-REACTION) dataset, which consists of more than 70 hours of audio and video data, from 2,222 subjects from the United States (1,138) and South Africa (1,084), aged from 18.5 – 49.0 years old.

The subjects within the dataset are reacting to a wide range of emotionally evocative stimuli (2,185 stimuli in total [18]). Each sample within the dataset has been self-annotated by the subjects themselves for the intensity of 7 emotional expressions in a range from 1-100: Adoration, Amusement, Anxiety, Disgust, Empathic Pain, Fear, Surprise.

The data is self-recorded via subjects’ own webcams in an environment of their choosing, including a wide variety of background, noise, and lighting conditions. Furthermore, different subjects spontaneously reacted with their faces and voices to varying degrees, such that the audio and multi-modal aspects of this sub-challenge will be particularly interesting to incorporate. The organisers also provide labels for detected (energy-based) vocalisations to aid participants in incorporating audio, with a total of 8,064 multi-modal recordings found to include vocalisations.

For the MuSe-Reaction sub-challenge the aim is to perform a multi-output regression from features extracted from the multi-modal (audio and video) data for the intensity of 7 emotional reaction classes. For this sub-challenge’s evaluation, the Pearson’s correlation coefficient ($\rho$) is reported as the primary baseline.

2.3 The MuSe-Stress Sub-challenge

The MuSe-Stress task is based on the multimodal Ulm-TSST database, for which subjects were recorded in a stress-inducing, free speech scenario, following the TSST protocol [31]. In the TSST, a job interview situation is simulated. Following a short period of preparation, a five-minute free speech oral presentation is given by the subjects. This presentation is supervised by two interviewers, who do not communicate with the subjects during the five minutes. Ulm-TSST comprises recordings of such TSST presentations of 69 participants (49 of them female), aged between 18 and 39 years. Overall, Ulm-TSST includes about 6 hours of data (cf. Table 1). On the one hand, the dataset features the audio, video, and text modalities. On the other hand, the physiological signals ECG, RESP, and BPM are provided. For extensive experiments on multimodal emotion recognition in TSST-based multimodal datasets see [9].

Ulm-TSST has been annotated by three raters continuously for the emotional dimensions of valence and arousal, at a 2 Hz sampling rate. Regarding valence, a gold standard is created by fusing the three corresponding annotator ratings, utilising the Rater Aligned Annotation Weighting (RAAW) method from the MuSe-Toolbox [45]. RAAW addresses the difficulties arising when emotion annotations – subjective in their nature – are to be combined into a gold standard signal. In short, RAAW first tackles the inherent rater lag by aligning the (per annotator) standardised signals via generalised Canonical Time Warping (CTW) [56]. After that, the Evaluator Weighted Estimator (EWE) [26] is applied to the aligned signals. EWE fuses the individual signals using a weighting based on each rater’s inter-rater agreement to the mean of all others. A detailed description of RAAW can be found in [45]. We obtain a mean inter-rater agreement of 0.204 ($\pm 0.200$) for valence.

As for the arousal gold standard, a different approach is employed. Instead of fusing the three annotators’ arousal ratings, we take the labels of last year’s MuSe-Physio sub-challenge as the arousal gold standard. Here, the annotator with lowest inter-rater agreement is discarded and replaced with the subject’s electrodermal activity signal (EDA) which is known to indicate emotional arousal [14]. This signal is downsampled to 2 Hz and smoothed using a Savitzky–Golay filtering approach (window size of 26 steps) in advance. Then, the two remaining annotators and the preprocessed EDA signal are again fused via RAAW, resulting in a mean inter-annotator agreement of 0.233 ($\pm 0.289$). This signal is called physiological arousal in the following. The motivation to employ this kind of gold standard is to obtain a more objective arousal signal. Considering such an objective criterion for arousal in addition to subjective annotations is especially relevant given the task at hand: in the job interview setting, individuals can be expected to try to hide their arousal, making it more difficult for annotators to recognise it. Detailed experiments on combining subjective annotations with objective physiological signals are provided in [8].
Ulm-TSST is split into train, development, and test partitions containing 41, 14, and 14 videos, respectively. The split is identical to the split used in last year’s challenge. Figure 1 shows the distributions of the valence and physiological arousal signals for the dataset.

2.4 Challenge Protocol
All challenge participants are required to complete the End User License Agreement (EULA) which is available on the MuSe 2022 homepage. Further, the participants must hold an academic affiliation. Each challenge contribution should be followed by a paper that describes the applied methods and provides the obtained results. The peer review process is double-blind. To obtain results on the test set, participants upload their predictions for unknown test labels on CodaLab. The number of prediction uploads depends on the sub-challenge: for MuSe-Humor and MuSe-Reaction, up to 5 prediction uploads can be submitted, while for MuSe-Stress, up to 20 prediction uploads are allowed. We want to stress that the organisers themselves do not participate as competitors in the challenge.

3 BASELINE FEATURES AND MODEL
To enable the participants to get started quickly, we provide a set of features extracted from each sub-challenge’s video data. More precisely, the provided features include up to five model-ready video, audio, and linguistic feature sets, depending on the sub-challenge. Regarding the label sampling rate, labels refer to 2 s windows in MuSe-Humor. The MuSe-Stress data is labelled at a 2 Hz rate. For MuSe-Reaction, there is one label vector of 7 classes per sample.

3.1 Pre-processing
All datasets are split into training, development, and test sets. For all partitions, ratings, speaker independence, and duration are taken into consideration (cf. Table 1). The videos in Passau-SFCH are cut to only include segments in which the respective coach is actually speaking. As the press conference setting can be seen as a dialogue between journalists and the coach, the answers given by each coach provide a natural segmentation of the Passau-SFCH data. For MuSe-Reaction – as can be seen in Table 1 –, a 60-20-20% split strategy is applied. There is no additional segmentation applied to clean the data further, each sample contains a single reaction to an emotional stimulus, and labels were normalised per sample to range from [0:1]. For further exploration, the participants are also provided with voice activity segments from the samples, which show to contain audio of substantial energy. In the Ulm-TSST dataset, we make sure to exclude scenes which are not a part of the TSST setting, e.g., the instructor speaking. Moreover, we cut segments in which TSST participants reveal their names. The Ulm-TSST dataset is not segmented any further.

3.2 Audio
All audio files are first normalised to -3 decibels and then converted from stereo to mono, at 16 kHz, 16 bit. Afterwards, we make use of the two well-established machine learning toolkits openSMILE [23] and DeepSpectrum [3] for expert-designed and deep feature extraction from the audio recordings. Both systems have proved valuable in audio-based Speech Emotion Recognition (SER) tasks [5, 10, 24].

3.2.1 eGeMAPS. The openSMILE toolkit [23] is used for the extraction of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [22]. This feature set which is proven valuable for SER tasks [7], also in past MuSe challenges (e.g., [50]), includes 88 acoustic features that can capture affective physiological changes in voice production. In MuSe-Humor, we use the default configuration to extract the 88 eGeMAPS functionals for each two second audio frame. For the audio of MuSe-Reaction, the 88 eGeMAPS functionals are extracted with a step size of 100 ms and window size of 1 second. Regarding MuSe-Stress, the functionals are obtained with a 2 Hz rate, using a window size of 5 seconds.

3.2.2 DeepSpectrum. The principle of DeepSpectrum [3] is to utilise pre-trained image Convolutional Neural Networks (CNNs) for the extraction of deep features from visual representations (e.g., Mel-spectrograms) of audio signals. The efficacy of DeepSpectrum features has been demonstrated for SER [39], sentiment analysis [2], and general audio processing tasks [4]. For our DeepSpectrum baseline experiments, we use DenseNet121 [30] pre-trained on ImageNet [40] as the CNN backbone. The audio is represented as a Mel-spectrogram with 128 bands employing the viridis colour mapping. Subsequently, the spectrogram representation is fed into DenseNet121, and the output of the last pooling layer is taken as a 1024-dimensional feature vector. The window size is set to one second, the hop-size to 500 ms.

3.3 Video
To extract specific image descriptors related to facial expressions, we make use of two CNN architectures: Multi-task Cascaded Convolutional Networks (MTCNN) and VGGFace 2. We also provide a set of Facial Action Units (FAUs) obtained from faces of individuals in the datasets. Further, participants are also given the set of extracted faces from the raw frames. In the videos of MuSe-Humor, typically more than one face is visible. As this sub-challenge’s objective is to predict the expression of humour of the coach, we only provide the faces of the respective coach and the features computed for them.

3.3.1 MTCNN. The MTCNN [54] model, pre-trained on the datasets WIDER FACE [53] and CelebA [35], is used to detect faces in the videos. Two steps are carried out to filter extracted faces that do not show the coach in Passau-SFCH: first, we automatically detect the respective coach’s faces using FaceNet embeddings of reference pictures showing the coach. The results of this procedure are then corrected manually. Ulm-TSST, in contrast, has a simple, static setting. The camera position is fixed and videos only show the TSST subjects who typically do not move much. Similarly,
for MuSe-Reaction, the video is captured from a fixed webcam. Hence, the performance of MTCNN is almost flawless for both Hume-Reaction and Ulm-TSST. The extracted faces then serve as inputs of the feature extractors VGGFace 2 and Py-Feat.

### 3.3.2 VGGFace 2
The purpose of VGGFace 2 is to compute general facial features for the previously extracted faces. VGGFace 2 [13] is a dataset for the task of face recognition. It contains 3.3 million faces of about 9,000 different persons. As the dataset is originally intended for supervised facial recognition purposes, models trained on it compute face encodings not directly related to emotion and sentiment. We use a ResNet50 [29] trained on VGGFace 2 and detach its classification layer, resulting in a 512-dimensional feature vector output referred to as VGGFace 2 in the following.

### 3.3.3 FAU
FAUs as originally proposed by Ekman and Friesen [21], are closely related to the expression of emotions. Hence, detecting FAUs is a promising and popular approach to the visual prediction of affect-related targets (e.g., [35]). We employ Py-Feat to obtain predictions for the presence of 20 different FAUs. We do not change Py-Feat’s default configuration, so that a pre-trained random forest model is used to predict the FAUs.

### 3.4 Language: Bert
In recent years, pre-trained Transformer language models account for state-of-the-art results in numerous Natural Language Processing tasks, also in tasks related to affect (e.g., [10]). In general, these models are pretrained in a self-supervised way utilising large amounts of text data. Subsequently, they can be fine-tuned for specific downstream tasks. For the transcripts of MuSe-Humor and MuSe-Stress, we employ a German version of the BERT (Bidirectional Encoder Representations from Transformers [19]) model11. No further fine-tuning is applied. For both Passau-SFCH and MuSe-Stress, we extract the BERT token embeddings. Additionally, we obtain 768 dimensional sentence embeddings for all texts in Passau-SFCH by using the encodings of BERT’s [CLS] token. In all cases, we average the embeddings provided by the last 4 layers of the BERT model, following [47].

### 3.5 Alignment
For each task, at least two different modalities are available. Typically, sampling rates per modality may differ. We sample the visual features with a rate of 2 Hz in all sub-challenges. The only exception is the FAUs in MuSe-Reaction, which are sampled at a 4 Hz rate. Regarding the audio features (DeepSpectrum and iGeMAPS), we apply the same frequency in MuSe-Humor and MuSe-Stress, while iGeMAPS features are obtained using a step size of 100 ms in MuSe-Reaction. As VGGish and FAUs are only meaningful if the respective frame actually includes a face, we impute frames without a face with zeros.

For MuSe-Humor, the binary humour label refers to frames of at most 2 seconds length. Hence, each label in MuSe-Humor corresponds to at most 4 facial and acoustic feature vectors. 2 Hz sentence embedding vectors are constructed by assigning every sentence to the 500 ms frames it corresponds to. If two sentences fall into the same frame, their embeddings are averaged to form the feature for that frame.

Regarding MuSe-Reaction, there is no alignment needed with labels, as each file is associated to a single vector of 7 emotional reaction labels.

For the MuSe-Stress sub-challenge, we provide label-aligned features. Hence, these features exactly align with the labels. We apply zero-padding to the frames, where the feature type is absent. Moreover, we downsample the biosignals in Ulm-TSST to 2 Hz, followed by a smoothing utilising a Savitzky-Golay filter. Participants are provided with both the raw signals and the downsampled ones.

In both Ulm-TSST and Passau-SFCH, manual transcripts are available. However, they lack timestamps. Hence, we reconstruct word level timestamps utilising the Montreal Forced Aligner (MFA) [37] tool. Here, we employ the German (Prosodylab) model and the German Prosodylab dictionary. The text features are then aligned to the 2 Hz label signal by repeating each word embedding throughout the determined interval of the corresponding word. In case a 500 ms frame comprises more than one word, we average over the word embeddings. Zero imputing is applied to parts where subjects do not speak. For the sentence embeddings in Passau-SFCH we choose an analogous approach, repeating and, if applicable, averaging the embeddings.

### 3.6 Baseline Model: LSTM-RNN
The sequential nature of the tasks makes recurrent neural networks (RNNs) a natural choice for a fairly simple baseline system. More specifically, we employ a Long Short-Term Memory (LSTM)-RNN. Initially, we train a single model on each of the available feature sets. Regarding MuSe-Stress, we separately train a model for both labels, valence and physiological arousal. We conduct an extensive hyperparameter search for each prediction target and feature. We thus optimise the number of RNN layers, the dimensionality of the LSTM’s hidden vectors and the learning rate. Of note, we also experiment with both unidirectional and bidirectional LSTMs. The code as well as the configurations found in the hyperparameter search are available in the baseline GitHub repository12.

Each label in MuSe-Humor is predicted based on all feature vectors belonging to the corresponding 2 s window. Hence, the sequence length in the MuSe-Humor training process is at most 4 steps.

In both MuSe-Reaction and MuSe-Stress, we make use of a segmentation approach which showed to improve results in previous works [43, 44, 47]. We find that a segmentation of the training data with a window size of 50 s (i.e., 200 steps) and a hop size of 25 s (i.e., 100 steps) leads to good results for MuSe-Stress. For MuSe-Reaction a slightly larger size of 500 steps and a hop size of 250, lead to more robust results.

Following the unimodal experiments, in order to combine different modalities, for MuSe-Humor and MuSe-Stress, we implement a simple late fusion approach. We apply the exact same training procedure as before, now treating the predictions of previously trained unimodal models as input features. In these experiments, we use one configuration per task, without performing a hyperparameter

---

1https://github.com/WeidiXie/Keras-VGGFace2-ResNet50
2https://py-feat.org
3https://huggingface.co/bert-base-german-cased
4https://github.com/WeidiXie/Keras-VGGFace2-ResNet50
5https://github.com/EIHW/MuSe2022
We apply the model described above for every sub-challenge. As this approach for late fusion is less suited to a multi-label strategy, we apply an early fusion strategy for MuSe-Reaction. For early fusion, we simply concatenate the best performing feature sets for each modality (audio and video), and then train a new model with the same hyperparameters from the uni-modal experiments. The code and configuration for the two fusion methods are also part of the baseline GitHub repository. Moreover, the repository also includes links to the best model weight files in order to ease reproducibility.

4 EXPERIMENTS AND BASELINE RESULTS

We apply the model described above for every sub-challenge. In what follows, we discuss the baseline results in more detail.

4.1 MuSe-Humor

The results for MuSe-Humor are given in Table 2. Each result is obtained from running the LSTM using the specified features with 5 different fixed seeds, consistent with the challenge setting.

Table 2: Results for MuSe-Humor. We report the AUC-Scores for the best among 5 fixed seeds, as well as the mean AUC-Scores over these seeds and the corresponding standard deviations.

| Features      | Development [AUC] | Test [AUC] |
|---------------|-------------------|------------|
| Audio         |                   |            |
| eGeMAPS       | .6861 (.6731 ± .0172) | .6952 (.6979 ± .0098) |
| DeepSpectrum  | .7149 (.7100 ± .0030) | .6547 (.6497 ± .0102) |
| Video         |                   |            |
| FAU           | .9071 (.9030 ± .0028) | .7960 (.7952 ± .0077) |
| VGGface 2     | .9253 (.9225 ± .0024) | **.8480** (.8412 ± .0027) |
| Text          |                   |            |
| BERT          | .8270 (.8216 ± .0045) | .7888 (.7905 ± .0035) |
| Late Fusion   |                   |            |
| A+T           | .8901 (.8895 ± .0005) | .7804 (.7843 ± .0037) |
| A+V           | .8252 (.8219 ± .0038) | .6643 (.6633 ± .0027) |
| T+V           | .8908 (.8893 ± .0015) | .8232 (.8212 ± .0017) |
| A+T+V         | .9033 (.9026 ± .0006) | .7973 (.7910 ± .0057) |

Evaluating audio and video features for the MuSe-Humor sub-challenge shows a clear pattern. The video-based features, FAU and VGGish, clearly outperform the audio-based features with VGGish accounting for an AUC of .8480 on the test set while eGeMAPS only achieves .6952 AUC. This comes as no surprise, given that the expression of humour is often accompanied by smile or laughter and thus recognisable from facial expressions. A manual inspection of the humorous segments confirms this intuition. Nevertheless, audio features are able to detect humour, too. Partly, this may be due to the presence of laughter. The performance of text features (.7888 on the test set) is slightly worse than for the features based on the video modality, but also better than the performance of the audio features. We find that the sentence-level BERT features outperform the token-level features. With the simple fusion of modalities, the performance is not improved. Specifically, the late fusion approach typically shows worse generalisation to the test data than the unimodal experiments. e.g., there is a discrepancy of about .16 between mean AUC on the development (.8219) and test (.6633) sets for the combination of audio and video.

4.2 MuSe-Reaction

Table 3 shows the results for the MuSe-Reaction baseline. As expected, the audio results are substantially lower than those from the video modality. Of particular note, as it pertains to audio, we see that the emotion-tailored feature-set of eGeMAPS performs poorly, almost 0.05 ρ lower on the development set than the DeepSpectrum features. Given that there is limited speech in the data set, this may be why the DeepSpectrum features perform better, as due to being spectrogram-based, they can potentially capture a more general acoustic scene and non-speech verbalisations potentially better.

For the video features, the FAUs are performing much better on the test set than VGGface 2 (although both are derived from faces), given the nature of the data being ‘reactions’, it may be that the facial action units are much more dynamic generally, and these features model more accurately the emotional expression occurring within the scene.

Interestingly, when we observe the individual class scores, we see that Amusement is consistently performing better than all other classes, a finding which is consistent for audio and video features (eGeMAPS: .148 ρ, and FAU: .405 ρ). As well as being the most likely class to contain non-verbal communication e.g., laughter, this performance may be due to the known ease of modelling highly aroused states of emotional expression [49]. However, it may also relate to the valence of the emotions as we can see from Figure 2, the Disgust class is the worst performing for FAU.

It is worth noting that in this case, the early-fusion of the two best-performing feature sets in each modality does not yield any beneficial results. This holds, although we do consider that through the use of a knowledge-based audio approach, we may see more improvement for audio, which may result in stronger performance via fusion.

4.3 MuSe-Stress

Table 4 reports the results obtained for MuSe-Stress. Consistent with results reported by some of last year’s participants [20, 27], the results for MuSe-Stress partly fail to generalise to the test data. With respect to single modality experiments, this observation is particularly significant for the video features. For example, the best seed for predicting physiological arousal based on Facial Action Units yields a CCC of .5191 on the development set, but only results in a CCC of .0785 on the test partition. The audio feature sets, in comparison, achieve better generalisation with the most extreme difference between development and test CCC being about .12 (for eGeMAPS on physiological arousal). Moreover, for both prediction targets, the DeepSpectrum audio features perform best among the unimodal approaches with CCC values of .4239 and .4931 on the test sets for physiological arousal and valence, respectively.
Table 3: Results for MuSe-Reaction. Reported is the mean Pearson’s Correlation Coefficient ($\rho$) for the 7 emotional reaction classes. For each feature and late fusion configuration, the result for the best of 5 fixed seeds is given. The respective mean and standard deviation of the results are provided in parentheses.

| Features            | Development | Test      |
|---------------------|-------------|-----------|
| Audio               |             |           |
| eGeMAPS             | .0583 (.0504 ± .0069) | .0552 (.0479 ± .0062) |
| DeepSpectrum        | .1087 (.0945 ± .0096) | .0741 (.0663 ± .0077) |
| Video               |             |           |
| FAU                 | .2840 (.2828 ± .0016) | .2801 (.2777 ± .0017) |
| VGGface 2           | .2488 (.2441 ± .0027) | .1830 (.1985 ± .0088) |
| Early Fusion        |             |           |
| A+V                 | .2382 (.2350 ± .0016) | .2029 (.2014 ± .0086) |

A surprising aspect of the unimodal results is that audio features yield better results for valence than for arousal, contrary to previous results in the domain of multimodal emotion recognition. For the visual features, no such tendency to work better for one of the two dimensions can be observed: FAUs lead to better results for predicting valence (mean CCC of .3878 on the test set) than for physiological arousal (.1135); the same is true for the VGGface 2 features (.1968 and .1576 mean CCC on the test set for valence and arousal, respectively). The textual BERT features account for higher CCCs on the development partition for valence (mean CCC of .3221) than for physiological arousal (.2828). Surprisingly, however, for arousal, they generalise better to the test data, while for valence, the mean BERT CCCs drops from .3221 to .1872 when evaluating on the test set. These partly counterintuitive results may be attributed to the job interview setting. Job interviewees typically suppress nervousness in an attempt to give a relaxed, sovereign impression. This might make the detection of arousal from audio and video difficult. The comparably stable performance of textual features for physiological arousal may be due to correlations between participants pausing their speech for a longer time – or hardly at all – and arousal. We find such correlations to exist for several participants.

We also experiment with the downsampled biosignals, motivated by some of last year’s approaches ([11, 34, 55]) to the task which used these signals as a feature. To do so, we concatenate the three signals (BPM, ECG, and respiratory rate) into a three-dimensional feature vector and normalise them. Here, severe generalisation and stability problems can be observed. To give an example, for arousal, the mean CCC performance of biosignal features on the development set is .2793, but for the test set, it drops to .1095. What is more, the standard deviations obtained with the biosignal results are consistently higher than those of any other modality. Because of these issues and in order not to inflate the number of experiments, we exclude the physiological modality from the late fusion experiments.

While valence prediction could not be improved by late fusion, the late fusion of the audio and text modality accounts for the best result on the test set for physiological arousal prediction (.4761 CCC), slightly surpassing the late fusion of audio and text (.4413) as well as DeepSpectrum (.4239). For valence, a generalisation issue for late fusion is apparent. To give an example, the late fusion of acoustic and visual features yields by far the best result on the development set (.6914) but only achieves a CCC of .4906 on the test set.

5 CONCLUSIONS
This baseline paper introduced MuSe 2022 – the 3rd Multimodal Sentiment Analysis challenge. MuSe 2022 features three multimodal datasets: Passau-SFCH with press conference recordings.
Table 4: Results for MuSE-Stress. Reported are the CCC values for valence, and physiological arousal. For each feature and late fusion configuration, the result for the best of 20 fixed seeds is given. The respective mean and standard deviation of the results are provided in parentheses. The combined results are the mean of arousal and valence test CCCs for each feature set.

| Features       | Development Arousal [CCC] | Test Arousal [CCC] | Development Valence [CCC] | Test Valence [CCC] | Combined [CCC] |
|----------------|--------------------------|-------------------|---------------------------|-------------------|----------------|
| Audio          |                          |                   |                           |                   |                |
| eGeMAPS        | 0.4112 (.3168 ± .0459)   | .2975 (.3338 ± .0836) | .5090 (.4744 ± .0244)     | .3988 (.3932 ± .0385) | .3482          |
| DeepSpectrum   | 0.419 (.3433 ± .0548)    | .4239 (.4372 ± .0323) | .5741 (.5395 ± .0207)     | .4931 (4826 ± .0324) | .4585          |
| Video          |                          |                   |                           |                   |                |
| FAU            | 0.5191 (.4257 ± .0475)   | .0785 (.1135 ± .0335) | .4751 (.3886 ± .0534)     | .2388 (.3878 ± .0560) | .1918          |
| VGGface 2      | 0.3171 (.2697 ± .0216)   | .2076 (.1576 ± .0285) | .2637 (.1106 ± .0739)     | .0936 (.1968 ± .1130) | .1506          |
| Text           |                          |                   |                           |                   |                |
| BERT           | 0.3280 (.2828 ± .0372)   | .3504 (.3218 ± .0423) | .3672 (.3221 ± .0285)     | .1864 (.1872 ± .0269) | .2683          |
| Physiological  |                          |                   |                           |                   |                |
| BPM + ECG + resp. | 0.3917 (.2793 ± .0782)   | .1095 (.1151 ± .0656) | .4361 (.2906 ± .0787)     | .1861 (.2141 ± .0953) | .1478          |
| Late Fusion    |                          |                   |                           |                   |                |
| A+T            | 0.4478 (.4409 ± .0038)   | .4761 (.4716 ± .0034) | .5243 (.4808 ± .0161)     | .3653 (.3163 ± .0211) | .4070          |
| A+V            | 0.5440 (.5167 ± .0142)   | .3777 (.4011 ± .0229) | .6914 (.6811 ± .0081)     | .4906 (.4969 ± .0184) | .4342          |
| T+V            | 0.4609 (.4425 ± .0112)   | .3303 (.3327 ± .0112) | .5144 (.4965 ± .0102)     | .2462 (.2364 ± .0082) | .2883          |
| A+T+V          | 0.5056 (.4940 ± .0070)   | .4413 (.4485 ± .0125) | .6104 (.5720 ± .0215)     | .3703 (.3455 ± .0258) | .4058          |

of football coaches annotated for humour, MuSe-Reaction containing emotional reactions to stimuli, and Ulm-TSST consisting of recordings of the stress-inducing TSST. The challenge offers three sub-challenges accounting for a wide range of different prediction targets: i) in MuSe-Humor, humour in press conferences is to be detected; ii) in MuSe-Reaction, the intensities of 7 emotion classes are to be predicted; and iii) MuSe-Stress is a regression task on the levels of continuous valence and arousal values in a stressful situation. Similar to previous iterations ([42, 43]), we employed open-source software to provide participants with an array of extracted features in order to facilitate fast development of novel methods. Based on these features, we set transparent and realistic baseline results. Features, code, and raw data are made publicly available. The official baselines on the test sets are as follows: .8480 AUC for MuSe-Humor as achieved using VGGface 2 features; a mean $\rho$ over all classes of .2801 for MuSe-Reaction is obtained utilising FAU, and a CCCs of 0.4761 and 0.4931 for physiological arousal and valence, respectively, for MuSe-Stress, based on DeepSpectrum features and a late fusion of audio and text modalities, respectively.

The provided baselines give a first impression on which features and modalities may be suited best for the different tasks. We believe that more refined methods of combining different modalities and features may lead to significant improvements over the reported baseline results. We hope that MuSe 2022 serves as a stimulating environment for developing and evaluating such novel approaches.

6 ACKNOWLEDGMENTS
This project has received funding from the Deutsche Forschungsgemeinschaft (DFG) under grant agreement No. 461420398, and the DFG’s Reinhart Koselleck project No. 442218748 (AUDiONOMOUS).
REFERENCES

[1] Shahin Amiriparian, Lukas Christ, Andreas König, Eva-Maria Meßner, Alan Cowen, Erik Cambria, and Björn W. Schuller. 2022. MuSe 2022: Baseline Paper MuSe’ 22, October 10, 2022, Lisboa, Portugal.

[2] Cong Cai, Yu He, Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Mingyu Xu, and Yekta Said Can, Bert Arnrich, and Cem Ersoy. 2019. Stress detection in daily life. In Proceedings of the 2nd International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2017) held in conjunction with the 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017). AAAI, IEEE, San Antonio, TX, 26–29.

[3] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self-assessment manikins. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2005. IEEE, IEEE, Kuala Lumpur, Malaysia, 1–5.

[4] Shahin Amiriparian, Tobias Hubner, Vincent Karas, Maurice Gerczuk, Sandra Ottl, and Bjorn W. Schuller. 2022. DeepSpectrumLite: A Power-Efficient Transfer Learning Framework for Embedded Speech and Audio Processing From Decentralized Data. Frontiers in Artificial Intelligence 5 (2022). 10. https://doi.org/10.3389/frai.2022.856232

[5] Alice Baird, Shahin Amiriparian, and Björn Schuller. 2019. Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation. In 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSIP). IEEE, IEEE, Kuala Lumpur, Malaysia, 1–5.

[6] Alice Baird, Lukas Stappen, Lukas Christ, Lea Schumann, Eva-Maria Meßner, and Björn W. Schuller. 2021. A Physiologically-adapted Gold Standard for Arousal During a Stress Induced Scenario. In Proceedings of the 2nd Multimodal Sentiment Analysis Challenge, co-located with the 29th ACM International Conference on Multimedia (ACMMM). ACM, Association for Computing Machinery, Changchun, China, 69–73.

[7] Alice Baird. Shahin Amiriparian, and Björn Schuller. 2019. Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation. In 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSIP). IEEE, IEEE, Kuala Lumpur, Malaysia, 1–5.

[8] Alice Baird, Andreas Triantafyllopoulos, Sandra Zankert, Sandra Ottl, Lukas Christ, Lukas Stappen, Julian Konzok, Sarah Sturmbauer, Eva-Maria Meßner, and Björn W. Schuller. 2021. An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress. Frontiers in Computer Science 3 (2021). 19. https://doi.org/10.3389/fcoms.2021.750284

[9] Björn W. Schuller and Anton Batliner and Christian Bergler and Cecilia Mascolo and Jang Jung and Andrea and Julia Leitao and Heysem Kayah and Shahin Amiriparian and Alice Baird and Lukas Stappen and Ottl and Maurice Gerczuk and Panagiotis Tzirakis and Chloe Brown and Jagmohan Chauhan and Andreas Grammenos and Apinan Hathasanombat and Dimitris Spithis and Tong Xia and Pietro Cicuta and Leon J. M. Rothkranz and Joeri Zwort and Jelle Creep and Casper Kaandorp. 2021. The INTERSPEECH 2021 Computational Paralanguage Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Pmetics. In Proceedings of the 2nd Multimodal Sentiment Analysis Challenge, co-located with the 29th ACM International Conference on Multimedia (ACMMM). ACM, Association for Computing Machinery, Changchun, China, 69–73.

[10] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H. Hellhammer. 1993. The use of electrodermal activity (EDA) measurement to understand consumer well-being: Development of the Humor Styles Questionnaire. In Proceedings of the 2nd Multimodal Sentiment Analysis Challenge. Association for Computational Linguistics, New York, NY, USA, 29–36.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[12] Anh-Quang Duong, Ngoc-Huyen Hu, Hyung-Jeong Yang, Gue-Sang Lee, and Soo-Hyun Kim. 2021. Multi-modal Stress Recognition Using Temporal Convolution and Recurrent Network with Positional Embedding. In Proceedings of the 2nd Multimodal Sentiment Analysis Challenge. Association for Computing Machinery, New York, NY, USA, 37–42.

[13] Alice Baird and Lukas Stappen and Sandra Ottl and Maurice Gerczuk and André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petru Lalaui, Shrikanth S Narayanan, et al. 2015. The Geneva minimalist acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 7 (2015). 190–202.

[14] Florian Eyben, Martin Wöllner, and Björn Schuller. 2010. Opensmile: the much faster and flexible open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia. Association for Computing Machinery, Firenze, Italy. 1459–1462.

[15] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, and Björn Schuller. 2022. EmoNet: A Transfer Learning Framework for Multi-Corpus Speech Emotion Recognition. In Proceedings of the 2nd Multimodal Sentiment Analysis Challenge. Association for Computing Machinery, New York, NY, USA, 305 – 315.

[16] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. Environment Psychological & Nonverbal Behavior. 19 (1978). 409–436.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[18] Adria Mallol-Ragolta, Nicholas Cummins, and Björn W Schuller. 2020. An In-Challenge EmoNet: A Transfer Learning Framework for Multi-Corpus Speech Emotion Recognition. In Proceedings of Affective Computing 33 (2022).

[19] Michael Grimm and Kristian Kroschel. 2017. Sentiment Analysis Using Image-based Deep Spectrum Features. In Proceedings of the 2nd Multimodal Sentiment Analysis Challenge. Association for Computational Linguistics, New York, NY, USA, 37–42.

[20] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. Ur-funny: A multimodal language and emotion bridged by continuous gradients. In Proceedings of the International Speech Communication Association, ISCA, ISCA, Stockholm, Sweden, 3512–3516.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2015. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Honolulu, Hawaii, 2261–2269.

[22] Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Models. ISCA, ISCA, Stockholm, Sweden, 3512–3516.

[23] Adria Mallol-Ragolta, Nicholas Cummins, and Björn W. Schuller. 2022. Hybrid Multimodal Fusion Encoders. In Proceedings of the 2nd Multimodal Sentiment Analysis Challenge. Association for Computational Linguistics, New York, NY, USA, 305 – 315.

[24] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. Environment Psychological & Nonverbal Behavior. 19 (1978). 409–436.
