The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress

Lukas Stappen
University of Augsburg
Augsburg, Germany

Alice Baird
University of Augsburg
Augsburg, Germany

Lukas Christ
University of Augsburg
Augsburg, Germany

Lea Schumann
University of Augsburg
Augsburg, Germany

Benjamin Sertolli
University of Augsburg
Augsburg, Germany

Eva-Maria Meßner
University of Ulm
Ulm, Germany

Erik Cambria
Nanyang Technological University
Singapore

Guoying Zhao
University of Oulu
Oulu, Finland

Björn W. Schuller
Imperial College London
London, United Kingdom

ABSTRACT

Multimodal **Se**ntiment Analysis (MuSe) 2021 is a challenge focusing on the tasks of sentiment and emotion, as well as physiological-emotion and emotion-based stress recognition through more comprehensively integrating the audio-visual, language, and biological signal modalities. The purpose of MuSe 2021 is to bring together communities from different disciplines; mainly, the audio-visual emotion recognition community (signal-based), the sentiment analysis community (symbol-based), and the health informatics community. We present four distinct sub-challenges: MuSe-Wilder and MuSe-Stress which focus on continuous emotion (valence and arousal) prediction; MuSe-Sent , in which participants recognise five classes each for valence and arousal; and MuSe-Physio , in which the novel aspect of 'physiological-emotion' is to be predicted. For this year's challenge, we utilise the MuSe-CaR dataset focusing on user-generated reviews and introduce the Ulm-TSST dataset, which displays people in stressful depositions. This paper also provides detail on the state-of-the-art feature sets extracted from these datasets for utilisation by our baseline model, a Long Short-Term Memory-Recurrent Neural Network. For each sub-challenge, a competitive baseline for participants is set; namely, on test, we report a Concordance Correlation Coefficient (CCC) of .4616 CCC for MuSe-Wilder ; .5088 CCC for MuSe-Stress , and .4908 CCC for MuSe-Physio . For MuSe-Sent an F1 score of 32.82 % is obtained.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Multimodal Sentiment Analysis; Affective Computing; Stress Detection; Electrodermal Activity; Multimodal Fusion; Challenge; Benchmark

ACM Reference Format:
Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. 2021. The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge (MuSe '21), October 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3475957.3484450

ACM ISBN 978-1-4503-8678-4
https://doi.org/10.1145/3475957.3484450

1 INTRODUCTION

In the 2nd edition of the **Mu**ltimodal **Se**ntiment Analysis in Real-life Media (MuSe) Challenge, we address four tasks incorporating novelties in each: emotion, physiological-emotion, and stress recognition as well as sentiment classification. In the *Multimodal Continuous Emotions in-the-wild sub-challenge* (**MuSe-Wilder**) and *Multimodal Sentiment in-the-wild Classification sub-challenge* (**MuSe-Sent**), one has to recognise emotional dimensions (arousal, valence) in a regression and classification manner. These tasks are based on work previously outlined for the MuSe 2020 challenge [42] and feature substantially improved methods for target creation. The first improvement is the application of *Rater Aligned Annotation Weighting* (RAAW), a gold standard fusion method for continuous annotations taking both the varied annotator reaction times (aligning) and inter-rater agreements (subjectivity) into account. Additionally, intelligent extraction of valuable features from continuous emotion gold-standards is used to cluster segment-level signals to representative classes so that contributors are faced with two five-way classifications of the level of valence and arousal. These two sub-challenges (MuSe-Wilder and MuSe-Sent) are motivated by the fundamental nature of gold-standard creation on which all tasks and applications of the field are premised. In the *Multimodal Emotional Stress sub-challenge* (**MuSe-Stress**), valence and arousal are predicted, from people in stressed dispositions. This sub-challenge is motivated by the high level of stress many people face in modern societies [6]. Given the increasing availability of low-resource equipment (e. g., smart-watches) able to record biological signals to

Table 1: Reported are the number (#) of unique videos, and the duration for each sub-challenge hh:mm:ss. Partitioning of the MuSe-CaR dataset is applied for each of the two sub-challenges. Ulm-TSST has a total duration of 5:47:27 after preprocessing, using the same split for MuSe-Stress and MuSe-Physio.

	MuSe-CaR			Ulm-TSST	
Partition	#	MuSe-Wilder	MuSe-Sent	#	Stress/ Psycho
Train	166	22:16:43	22:35:55	41	3:25:56
Devel.	62	06:48:58	06:49:46	14	1:10:50
Test	64	06:02:20	06:14:08	14	1:10:41
$\sum$	291	35:08:01	35:39:49	69	5:47:27

track wellbeing, we propose the *Multimodal Physiological-Arousal sub-challenge* (**MuSe-Physio**). Adapted from MuSe-Stress, the arousal annotations from humans are fused (using RAAW) with galvanic skin response (also known as Electrodermal Activity (EDA)) signals for predicting physiological-arousal. Both are set up as regression tasks offering additional biological signals (e. g., heart rate, and respiration) for modelling.

For the introduced sub-challenges, two datasets are utilised. As last year [42], we reuse the Multimodal Sentiment Analysis in Car Reviews data (**MuSe-CaR**) [43] for the MuSe-Wilder and MuSe-Sent sub-challenges. Including almost 40 hours of video data, it is the most extensive emotion annotated multimodal dataset, gathered in-the-wild with the intention of further understanding real-world Multimodal Sentiment Analysis (MSA), in particular the emotional engagement that takes place during English-speaking product reviews. Within MuSe-CaR, the subjects are aged between 20 and 60 years, and the spoken word is entirely transcribed. For the first time, a sub-set of the novel audio-visual-text Ulm-Trier Social Stress dataset (**Ulm-TSST**), featuring German-speaking individuals in a stress-induced situation caused by the Trier Social Stress Test (TSST), is used in this year's MuSe-Stress and MuSe-Physio sub-challenges. The initial state of Ulm-TSST consists of 110 individuals (10 hours), richly annotated by self-reported, and continuous dimensional ratings of emotion (valence and arousal). In addition to audio, video, textual features, the Ulm-TSST includes four biological signals captured at a sampling rate of 1 kHz; EDA, Electrocardiogram (ECG), Respiration (RESP), and heart rate (BPM). Both datasets provide a common testing bed with a held-back labelled test set, to explore the modalities and employ state-of-the-art models under well-defined and strictly comparable conditions.

The goal of the MuSe challenges are to provide a paradigm that is of interest across several communities and to encourage a fusion of disciplines. We ideally aim for participation that strives for the development of unified approaches applicable to what we perceive as synergistic tasks which have arisen from different academic traditions: on the one hand, we have complex, dimensional emotion annotations that reflect a broad variety of emotions, grounded in the psychological and social sciences relating to the expression of behaviour, and on the other hand, we provide sentiment classes as it is common in sentiment analysis from (multimodal) text-focused modelling. These fields are rooted within Affective Computing (AC), of which a core aspect is the intelligent processing of uni-modal signals. Up to now, the focus in AC when predicting emotion such

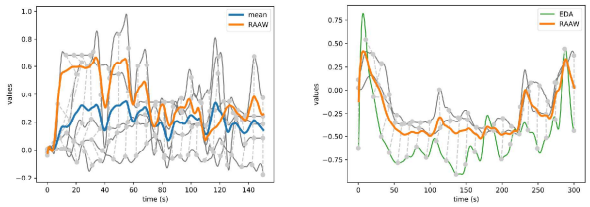

Figure 1: Sample gold standard computation using RAAW for emotion annotations (dark grey), aligned (warping paths in light grey), and fused from the MuSe-Toolbox [49]. The instance on the left side is from MuSe-Wilder (id: 100), the resulting signal (orange) is displayed and compared to the average annotation (blue). The right image displays a fusion from MuSe-Physio (id:11), where raters 1 and 2 for arousal (dark grey) are combined with the EDA signal (green).

as by valence and arousal dimensions, was mostly with lower attention to research made on textual information [21, 39]. However, the communities appear to be converging even more in recent years (such as supported by the MuSe 2020 [42] challenge), finding great benefit from multimodal approaches [2, 15, 33]. As an example, both the 2020 and 2021 INTERSPEECH Computational Paralinguistics (ComParE) Challenge have included textual features in an endeavour to more reliably predict valence [37, 38]. The second motivation of MuSe is to compare the merits of each of the core modalities (audio, visual, biological, social, and textual signal), as well as various multimodal fusion approaches. Participants can extract their own features or use the provided standard feature sets from the baseline models.

The paper's structure is as follows: First, the four sub-challenges with the corresponding datasets are explained in detail, followed by a description of the challenge conditions. Next, we describe the extracted features from different modalities and the applied pre-processing and alignment for the baseline modelling. Finally, we summarise our baseline results and conclude our findings. A summary of the challenge results can be found in [46].

2 THE FOUR SUB-CHALLENGES

In the following, we describe and highlight the aforementioned novelties of each sub-challenge, as well as include the guidelines for participation. The evaluation metric for all continuous time-based regression tasks is Concordance Correlation Coefficient (CCC), a well-understood measure [29] of reproducibility, often used in challenges [35, 42, 52]. The classification task (MuSe-Sent) is evaluated in F1 score (macro), a measure robust to class-imbalance. For all challenges with more than one target, the mean of all measures is taken for the final performance evaluation.

2.1 The MuSe-Wilder Sub-Challenge

The MuSe-Wilder is an extension of the MuSe-Wild 2020 sub-challenge, where participants had to predict emotional dimensions (valence, arousal) in a time-continuous manner. The amount of data utilised from MuSe-CaR is shown in Table 1. The valence dimension is often referred to as the emotional component of the

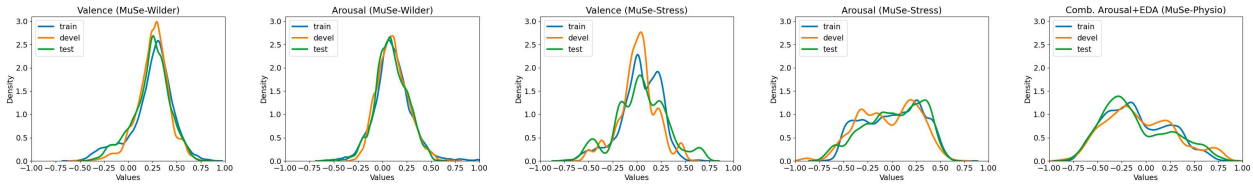

Figure 2: Frequency distribution in the partitions train, (devel)opment, and test for the continuous prediction sub-challenges MuSe-Wilder, MuSe-Stress, and MuSe-Physio. For all sub-challenges, the value distributions between the partitions are fairly similar.

generic term sentiment analysis and is often used interchangeably [26, 32, 51]. Human annotation of continuous emotions leads to disagreements between raters, e. g., due to differences in perception [19] and reaction time [27], which should be mitigated by fusion to a gold standard. Since this signal is the prediction target, a variety of fusion methods are available in the literature [16, 28] and this development has motivated other challenges [34].

This year's MuSe-Wilder emotion recognition task is based on a completely novel continuous annotator fusion technique RAAW, which targets the difficulties of combining subjective emotion annotations for a gold standard annotation present. For this, we employ our fusion method on a minimum of five different ratings that weights inter-rater agreements as well as considers the varied reaction times as displayed in Figure 1. The varying rater lag that is inherent to all annotation signals will be targeted by aligning the standardised (per annotator) ratings using a generalised Canonical Time Warping (CTW) method [57]. The Evaluator Weighted Estimator (EWE) [16] is then used to fuse the aligned, individual signals by weighting a signal depending on the inter-rater agreement to the mean of all others. This technique is described in length in [49]. The resulting distribution is shown in Figure 2.

2.2 The MuSe-Sent Sub-challenge

Mapping continuous emotion annotations to discrete classes are considered a highly ambiguous and challenging task and have so far hardly been computed successfully in a time-continuous fashion [53]. In general, classes are often considered a simplified concept for interpretation compared to dimensional representations. In MuSe-Sent, participants will have to predict five advanced sentiment classes for each emotion dimension of valence and arousal on a segment-level, based on audio-visual recordings and the transcribed speech of MuSe-CaR. The sub-challenge uses the topic-based segmentation from MuSe 2020 [42]. The classes are extracted using a novel method of the MuSe-Toolbox [49]. which aims to find a mapping between continuous dimensional and categorical representations of emotion through the extraction of time-series features and the application of unsupervised clustering.

More specifically, we first extract a range of time-series features on a segment-level[1] based on the continuous RAAW-fused annotations. The absolute features are normalised depending on the varying length of a segment to limit undesirable properties solely due to the influence of the segment length. To reduce the feature space, we apply Principal Component Analysis (PCA) to project our data to a five dimensional space of principal components which are derived from the eigenvectors of the covariance matrix. The transformed data is further clustered into five class clusters using a) the k-means algorithm [24] for valence and b) a Gaussian Mixture clustering model [10] for arousal. To ensure that the development and test set have no effect on the generated classes, we only apply this process on the training set segments. The segments belonging to the development and test partitions, are then 'predicted' by assigning the cluster with the closest centre to each data-point. These clusters are evaluated through both qualitative and quantitative measures: (1) the amount of data-points of the smallest class is larger than a quarter of by-chance-level[2] (2) to evaluate cluster cohesion and separation, the widely used Silhouette Coefficient (SILC) [36] is calculated, ranging from -1 to 1 (closer to 1 is superior). For the two chosen settings, we achieve a SILC of 0.19 and 0.10, respectively for valence and arousal clusters. The PCA leads to a denser representation along the orthogonal axes, making a higher SILC value hard to achieve, since the metric is prone to error when clusters show different kinds of cluster densities [22], which naturally occurs in this setting.

Since the features reflect characteristics of the emotional annotation and not just the mean value as in last years' MuSe-Topic task [42], class descriptions, i. e., low, medium, or high would inadequately reflect the meaning. With this in mind, to gain understanding of the classes, we display the most distinctive features in Figure 3 for interpretation showing the named valence classes as $V_{\#}$ and arousal as $A_{\#}$ while # represents the class number, not implying any specific order. For example, segments from classes V_1 and A_2 have a comparatively large (to the mean of all other classes) 'standard deviation' and 'sum of changes', which indicates a higher annotation fluctuation and intensity than other classes. The distribution of segments across the classes can be found in Table 2.

2.3 The MuSe-Stress Sub-challenge

In the MuSe-Stress, participants will have to predict valence and arousal in a time-continuous manner. This sub-challenge is motivated by real-world applications for emotion recognition and further motivated by stress in modern life. In this novel sub-challenge,

[1] arousal: median, standard deviation, percentile {10, 90}, relative energy, relative sum of changes, relative number of peaks, relative longest strike {below, above} mean, and relative count below mean; Valence: the same features as for arousal, and additionally: mean, percentile {5, 25, 33, 66, 75, 95}, and the percentage of reoccurring data-points to all data-points

[2] For example, five classes resemble a by-chance level of 20 %, thus, the smallest class have to cover at least 5 % of the data points

Table 2: Distribution of the valence and arousal classes across partitions used in the MuSe-Sent sub-challenge as a result of our configured class search from the MuSe-Toolbox [49].

	Valence				Arousal		
	Train.	Devel.	Test		Train.	Devel.	Test
0	528	71	89	0	612	249	178
1	552	159	277	1	534	135	194
2	1178	458	378	2	312	96	53
3	1112	405	271	3	1255	388	448
4	837	242	245	4	1494	467	387
$\sum$	4207	1335	1260	$\sum$	4207	1335	1260

the idea of 'multimodal' sentiment analysis is pushed further by the inclusion of biological signals that have been shown to be applicable for recognising physiological stress [31], and for emotion recognition [40].

Participants are provided with the multimodal Ulm-TSST database, in which subjects were recorded under a highly stress-induced free speech task, following the TSST protocol [20]. For the TSST, after a brief period of preparation the subjects are asked to give an oral presentation, within a job-interview setting, observed by two interviewees who remain silent for the period of five minutes. To allow consistent data partitions, we only keep data recorded under the same experimental conditions. The resulting 69 participants (49 of them female) are aged between 18 and 39 years, providing a total amount of about 6 hours of data for the MuSe-Stress and MuSe-Physio sub-challenges (cf. Table 1). Besides audio, video, and text, the participants can optionally utilise the ECG, RESP and BPM signals.

The dataset has been rated by three annotators continuously for the emotional dimensions of valence and arousal, at a 2Hz sampling rate, and a gold standard is obtained by the fusion of annotator ratings, utilising the RAAW method, as described in Section 2.1 from the MuSe-Toolbox [49]. When creating the fusion a mean CC inter-rater agreement of 0.204 (± 0.200) for valence and 0.186 (± 0.230) for arousal is obtained. The distributions of the valence and arousal signals for the dataset are depicted in Figure 2.

2.4 The MuSe-Physio Sub-challenge

In the cross-modal MuSe-Physio , participants will have to predict a combined signal of arousal and EDA. Again, for this task, the Ulm-TSST dataset is employed, where the TSST was utilised as a standardised and renowned stressor, allowing for a controlled setting with high-quality data while maintaining a naturalistic subject behaviour.

Physiological signals, including EDA have been used as a feature in at least one other multimodal emotion challenge [9]. However, we consider this sub-challenge to be the first time the physiological signal is combined with the emotional – human-annotated – signal. From all the biological signals available in the Ulm-TSST dataset, we choose to use the EDA signal, as not only are the signal characteristics subjectively similar to continuous emotion as exemplarily depicted in Figure 1, but the signal itself has been shown in the literature to be a psycho-physiological indication of emotional arousal [7]. Given that in the context of an interview, arousal may also appear to be a more hidden emotion, we consider that the fusion of arousal and EDA may improve recognition and offers a more objective marker for a speaker's arousal [7]. Further variants are introduced in [4]. To obtain the combined emotion and EDA signal gold standard, we again utilise the RAAW fusion strategy. However, in this case, the lowest weighted annotator is excluded and replaced with the EDA signal. The EDA signal is downsampled to 2 Hz and smoothed slightly before fusion through a Savitzky–Golay filtering approach (window size of 26 steps), to avoid irrelevant, fine-grained artefacts in the signal. For this gold standard of emotion, we obtain an inter agreement of 0.233 (±0.289), which was improved compared to the arousal gold standard obtained in MuSe-Stress .

2.5 Challenge Protocol

As part of the mandatory prerequisites required to play a part in the challenge, interested participants are obliged to download and fill in the End User License Agreement (EULA) which is put forward through the homepage[3]. On top of this, participants are further required to hold an academic affiliation. Each participation must be accompanied by a paper (6-8 pages in length including references) reporting the results obtained and methods applied. The organisers also consider general contributions in the field. Peer review is double-blind. To obtain results on the test set, the participants can upload their predictions up to five times per sub-challenge, whose labels are unknown to them. We want to point out that the organisers only evaluate the participants' results but do not participate themselves as competitors in the challenge.

3 BASELINE FEATURES AND MODEL

To save effort and time which would be incurred by the participants while extracting various features from the large datasets provided, we put forth a selection of features drawn from the video data for each sub-challenge. In a more elaborate outline, the available features comprise of seven model-ready video, audio, and linguistic feature sets[4]. The amalgamation of features provided surpasses most other related audio-visual challenges [9, 11, 55]. In respect to the annotation sampling rate, the features are extracted at a step size of 0.25 s for the MuSe-CaR and 0.5 s for the Ulm-TSST dataset.

3.1 Pre-processing

The data of both datasets has been partitioned into a Train, Development, and Test partition. Emotional ratings, speaker independence, and duration are considered when creating the partitions (cf. Table 1 for an overview). Since the amount of recordings made available between sub-challenges can vary, so too does the time required to extract the most applicable features during the pre-processing stages. Aiming to minimise the distortion of the task objectives, we deliberately omit advertisement sections of the videos for the MuSe-CaR-based sub-challenges. In the Ulm-TSST dataset, each video is cut to exclude scenes outside of the TSST setting, e. g.,

[3]https://www.muse-challenge.org/muse2021

[4]Note: Furthermore, we place it at the applicants' disposal to use (unaligned) features from MuSe2020 for MuSe-Wilder and MuSe-Sent as well as external datasets and features under the condition that this should be clearly explained in their accompanying paper. These sources could be, e. g., commercial or academic feature extractors, libraries, or pre-trained networks.

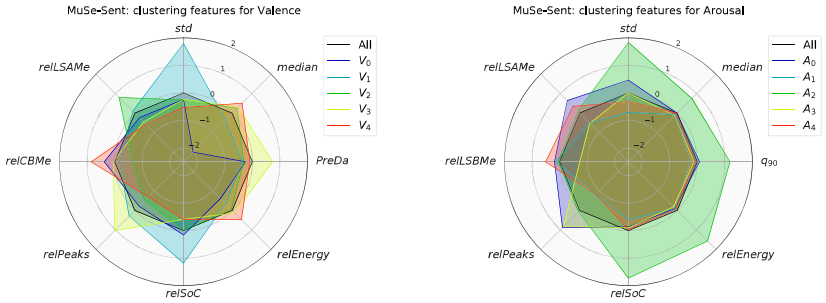

Figure 3: Mean of selected clustering features for each of the created classes which are used in the MuSe-Sent sub-challenge. The features shown are standard deviation (*std*), *median*, the 90th percentile (q_{90}), percentage of reoccurring datapoints to all datapoints (*PreDa*), relative energy (*relEnergy*), relative sum of changes (*relSoC*), relative number of peaks (*relPeaks*), relative count below mean (*relCBMe*), relative longest strike below mean (*relLSBMe*), and relative longest strike above mean (*relLSAMe*). Features indicated by "relative" (*rel*) are normalised by segment length. Additionally, all features are standardised, hence, the mean value of all data is always equal to zero. Illustrations generated by the MuSe-Toolbox [49]

excluding participants' names. For both datasets, the segments are crafted with the focus on the active voice based on the sentence transcriptions or if a visible face applies. For MuSe-Sent , we adjacent segments in instances where the segments deals with the same topic and the gap is less than two seconds.

3.2 Acoustic

openSMILE and DeepSpectrum are well-established tools for the extraction of acoustic emotional feature representations. Most notably, they have proved valuable in the extraction of audio processing tasks in renowned challenges in speech emotion recognition (SER) [37, 38]. For all acoustic features, a six second window size is applied. In the first step of the pre-processing pipeline, the full audio is extracted from a given video. The second step is the conversion of the audio from stereo to mono to 16 kHz, 16 bit after its normalisation to -3 decibels.

3.2.1 eGeMAPS. The prevalent open-source openSMILE toolkit [13] is used to extract the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [12]. Comprising of 88 acoustic parameters for automatic voice analysis tasks [45], it is a minimal set of handcrafted features relying on affective physiological changes in voice production that has previously proven valuable for a variety of emotion research [3, 42, 44].

3.2.2 DeepSpectrum. The prime function of DeepSpectrum [1] is to utilise the spectral features acquired from speech instances within a pre-trained image recognition Convolutional Neural Networks (CNNs). The consecutive inputs result in the extraction of feature vectors. A commonly applied architecture in this framework is VGG-19 [41]. Here, we keep the default settings for extraction to obtain a 4 096 dimensional feature set.

3.2.3 VGGish. In addition, we extract VGGish functions [18] pretrained on an extensive YouTube audio dataset (AudioSet) [14]. The underlying data contains 600 classes, and the recordings contain a variety of 'in-the-wild' noises that we expect to be beneficial to obtain robust features from our 'in-the-wild' videos. By aligning the frame and hop size to the annotation sample rate, we extract a 128-dimensional VGGish embedding vector every 0.25 s from the underlying log spectrograms.

3.3 Vision

Extracting specific image descriptors that match certain attributes, e. g., face, remains the paramount focus of most visual feature extractors. Our offered visual feature sets are inclined to capture the entire surroundings as well as analysing human behaviour synthesised from gesture and facial expressions. Participants are also provided with an array of extracted faces which are directly extracted from the raw frames.

3.3.1 MTCNN. The MTCNN [56] is used to distinguish facial expressions captured in the videos, pretrained on the data sets WIDER FACE [54] and CelebA [23]. For MuSe-CaR, we examined the extraction as described in detail in [42], where an F1 score of 86 % on a labelled subset was achieved. Compared to these highly dynamic camera positionings (zoom, free etc.), Ulm-TSST has a static setting. In an visual inspection aimed to control the performance, an apparently flawless extraction was found. The extractions were ultimately put in use as inputs for VGGFace and OpenFace .

3.3.2 VGGFace. VGGFace (version 1) [30] is aimed at the extraction of general facial features for images obtained by MTCNN in cropped versions. The visual geometry group of Oxford introduced the deep CNN referred to as VGG16 [41]. The training data constitutes of 2.6 million faces and over 2 500 identities. The VGGFace architecture was originally intended for supervised facial recognition purposes [41]. However, detaching the top-layer of a pretrained version results in a 512 feature vector output referred to as VGGFace. Presenting high levels of performance while consuming less data is the main advantage held for VGGFace in comparison to other facial recognition models.

3.3.3 OpenFace. Facial features in 2D (136 dimensions) and 3D (204 dimensions), gaze positions (288 dimensions), intensity and activity of 17 Facial Action Units (FAUs) for both center and left side, and 6 head stances were extracted from cropped faces identified using MTCNN . This was achieved through the wide array of facial

features offered by the OpenFace [5] toolkit. For the Ulm-TSSTdata challenge, we only provide intensity, as activity features appear to be of less use for this task.

3.3.4 Xception. Generally used to extrapolate generic vision features, Xception [17] should provide participants with environmental features using stacked residual blocks[5]. Among other challenges, it came in first on the ILSVRC 2015 classification challenge. The network is pre-trained on the ImageNet dataset compromising of 350 million images and 17 000 classes. The then frozen network architecture prepossesses a given frame through the layers until the last fully connected layer from which a 2 048 deep feature dimensional vector is obtained.

3.4 Language: Bert

The text feature extraction process employs a Transformer language model, namely Bidirectional Encoder Representations from Transformers (BERT) [8], which have already been successfully used for a variety of NLP tasks [37, 38, 44, 47, 48]. BERT pre-trains its deep representations on context of unlabelled text before fine-tuning them on a broad selection of down-streaming NLP tasks. During inference, the context-based representations are preserved, excerpting one vector per word. This is in contrast to static word embeddings which give one vector per word independent of the context. Our features are the sum of the last four BERT layers resulting in a 768 dimensional feature vector analogous to [50]. For MuSe-Wilder and MuSe-Sent, the base variant of BERT, pretrained on English texts, is used. Analogously, as the Ulm-TSST data set is in German, for MuSe-Stress and MuSe-Physio, the BERT (base) pretrained on German texts is utilised.

3.5 Alignment

The extensive assortment of features are from three modalities. The corresponding sampling rate of each modality differs, which leads to a different length of the extracted features along the time axis. All visual features are incessant through the video with a frame sampling of 4 Hz for MuSe-CaR and 2 Hz for Ulm-TSST, which is equivalent to the labelling rate. The audio sampling of DeepSpectrum, and that of eGeMAPS apply the same frequency. VGGish and Facial Action Units are the only feature sets relying only on frames where a face is observable. By the nature of text, the corresponding features do not follow a fixed sampling rate, as the duration of a spoken word varies.

For each sub-challenge, we make label-aligned features available. These have accurately the same stretch and time-stamps as the provided label files. We apply zero-padding to the frames, where the feature type is absent. Such instances include OpenFace, when no face appears or extraction fails, e. g., when only small faces appear in the original frame. The text features are repeated for the interval of a word and non-linguistic parts are also imputed with zero vectors. MuSe-CaR offers automatic, word-aligned transcriptions [43]. For Ulm-TSST, manual transcripts of the videos are available. We use the Montreal Forced Aligner (MFA) [25] tool to obtain time-stamps on the word level. The MFA includes pretrained acoustic models, grapheme-to-phoneme models, and pronunciation dictionaries for various languages. We use the German (Prosodylab) model and the German Prosodylab dictionary to align the Ulm-TSST transcripts. The time-stamps yielded by the MFA are used to align the word embeddings to the 2 Hz frames in the Ulm-TSST dataset.

3.6 Baseline Model: LSTM-RNN

In order to address the sequential nature of the input features, we utilise a Long Short-Term Memory (LSTM)-RNN based architecture. The input feature sequences are input into uni- and bi-directional LSTM-RNNs with a hidden state dimensionality of $h = \{32, 64, 128\}$, to encode the feature vector sequences. We test different numbers of LSTM-RNN layers $n = \{1, 2, 4\}$. Based on experiences from initial experiments, some hyperparameter searches are task-dependently executed: MuSe-Wilder we search for a suitable learning rate $lr = \{0.0001, 0.001, 0.005\}$; for *MuSe-Sent* $lr = \{0.001, 0.005, 0.01\}$; for *MuSe-Stress* and *MuSe-Physio* $lr = \{0.0001, 0.0002, 0.0005, 0.001\}$. As we observed overfitting in some settings of MuSe-Physio, we also tried L2-Regularisation with a penalty of 0.01 for this task.

The sequence of hidden vectors from the final LSTM-RNN layer is further encoded by a feed-forward layer that outputs either a one-dimensional prediction sequence of logits for each time step (regression), or a single-value per prediction target (classification).

In the training processes, the features and labels of every input video are further segmented via a windowing approach [42, 43, 50]. For MuSe-Wilder and MuSe-Sent, we use a window size of 200 steps (50 s) and a hop size of 100 steps (25 s). For MuSe-Stress and MuSe-Physio, a window size of 300 steps (150 s) and a hop size of 50 steps (25 s) proved to be reasonable choices.

3.7 Fusion

We apply decision-level (late) fusion to evaluate co-dependencies of the modalities. The experiments are restricted to the best performing features from each modality only. For decision-level fusion, separate models are trained individually for each modality. The predictions of these are fused by training an additional LSTM-RNN model as described above. For all continuous regression tasks, we apply uni-directional version with $lr = 0.0001$, $h = 64$, and $n = 1$, and for MuSe-Sent a bi-directional one with $lr = 0.005$, $h = 32$, and $n = 2$.

4 EXPERIMENTS AND BASELINE RESULTS

For all sub-challenges, the same network architecture is applied (cf. Section 3.6). For reproducibility, we provide the detailed set of hyperparameters for our best models for each experiment, alongside our code in the corresponding GitHub repository[6], where also a link to the fully trained model weights can be found. In the following section, we give an overview of all baseline results as summarised in Table 3.

4.1 MuSe-Wilder

We evaluated several feature sets and combinations for the prediction of the continuous valence and arousal (cf. Table 3). The input features BERT in combination with our baseline architecture set

[5] not used for MuSe-Stress or MuSe-Physio as recording environment for Ulm-TSST changes only minimally and participants showed minimal movement due to their stressful situation.

[6] https://github.com/lstappen/MuSe2021

to $lr = 0.005$, $h = 128$, and $n = 4$ show superior results for the prediction of valence leading to a CCC of .4613 on the development and .5671 CCC on test set. For the prediction of arousal, using DeepSpectrumas input features and setting $lr = 0.001$, $h = 64$, and $n = 2$, yields the best result of all applied systems with a CCC of .3386 on the test set. Generally, we found that a unidirectional LSTM-RNN achieves better results for this task than complex bidirectional configurations and is used for the reported MuSe-Wilder results. When fusing the best performing features of all three modalities DeepSpectrum, VGGface, and BERT , the late fusion technique reaches .4863 and .5974 for valence and .4929 and .3257 for arousal on the development and test set, respectively. This technique yields the highest combined metric (mean of valence and arousal) of .4616 (on test) and is our baseline.

4.2 MuSe-Sent

For the classification tasks in the MuSe-Sent sub-challenge, we give an overview in Table 3 and further provide the confusion matrices for the best uni-modal setups tested on valence and arousal in Figure 4. For the prediction of valence on uni-modal feature inputs, the best result is achieved using the text-based BERT features as input and a baseline model setting of $lr = 0.001$, $h = 64$ and $n = 4$ (bi-directional), with an F1 score of 32.68 % on the development and 31.90 % on the test set. Using the audio-based DeepSpectrum features with a $lr = 0.001$, $h = 128$, and $n = 2$ (bi-directional), results in our highest F1 score for arousal with 33.52 % on the development and 33.16 % on the test set. Across both targets, we find that LSTM-RNN models with a bidirectional setting and at least two layers tend to achieve better results for this task than smaller architectures. Partially, we see improvements when we apply late fusion. For valence, utilising the predictions of VGGface and BERT yields a performance of 32.91 % F1-score on the test set. For arousal, the audio-visual fusion set-up (VGGish and FAU) also improves on the test set, with an F1 score of 35.12 %. Looking at the combined scores (mean of valence and arousal), using the BERT features alone comes out on top for the development set, reaching a 35.48 % F1 score, while fusing the video- and text-based predictions achieves the highest F1 score of 32.82 % on the test set.

4.3 MuSe-Stress

The best results from all feature sets and fusion of modalities are reported in Table 3[7]. Having searched the hyperparameter combinations mentioned, we achieve the best results on all settings with a 4-layered unidirectional LSTM equipped with 64-dimensional hidden states and a learning rate of 0.0002 with a maximum of 100 epochs, and early stopping with a patience of 15 epochs. Here, eGeMAPS outperforms all other single feature sets for the prediction of valence, achieving .5845 CCC on development and .5018 CCC on the test set. Regarding arousal, eGeMAPS is the best scoring single feature set, leading to .4304 and .4416 CCC on development and test set, respectively. For both valence and arousal prediction, the fusions of the best audio and vision feature sets result in the best performance overall. They achieve CCC values of .6966 (development) and .5614 (test) for valence, and .5043 (development) and .4562 (test) for arousal.. It is notable that the text feature set, BERT , performs considerably worse than the best audio and visual features.

[7] Of note, besides eGeMAPS, we also normalise the VGGish features for predicting arousal.

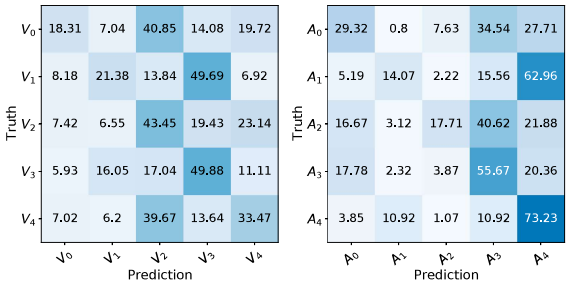

Figure 4: Relative confusion matrices over the 5 valence (left) and arousal (right) classes on the development partition for the MuSe-Sent sub-challenge. The results were achieved with the LSTM baseline model using the eGeMAPSfeature set with hyperparameters of $n = 4$ (bi-directional), $h = 128$, and a $lr = 0.001$ for valence, and for arousal the BERT features with a uni-directional model setting of $n = 2$, $h = 64$ and a $lr = 0.01$.

We found that, in general, valence reaches a stronger final result than arousal for this task. While this is not surprising for text features, it counters conventional expectations for the audio modality. A major reason for the poor arousal prediction results may be the TSST scenario, which imitates a job interview. Typically, interviewees try to remain neutral, suppressing nervousness, hence, the arousal shown to their counterpart would be minimal, thus, making arousal more difficult to detect in the Ulm-TSST data set than other comparable multimodal emotion recognition data sets.

Although we do not evaluate the provided bio-signal features systematically here, we encourage participants to explore them. To give an example, we achieve .2495 and .1537 CCC for valence on the development and test sets, respectively, by using only the three provided bio-signals (at a sampling rate of 2 Hz) as features in a four-layer LSTM. Similarly, they show also promising results for the prediction of arousal, reaching .1954 CCC on the development and .2189 CCC on test partition.

4.4 MuSe-Physio

For MuSe-Physio , the same LSTM configuration as for MuSe-Stress is applied. The results are reported in Table 3. Again, audio and video features considerably outperform the textual BERT features. While BERT only achieves .2583 and .1604 CCC on development and test data, respectively, the best audio feature set (DeepSpectrum) leads to .4423 and .4162 CCC on development and test data, respectively. Consistently, visual features outperform the textual ones, too. The best visual feature set (VGGface) yields .3903 and .4582 on development and test data respectively and hence shows comparable performance to DeepSpectrum. Like in MuSe-Stress , the late fusion of the best audio (VGGish) and video (VGGface) predictions yield the best results, namely .4913 CCC on development data and .4908 CCC on test data.

Table 3: Reporting Valence, Arousal, Combined ($0.5 \cdot Arousal + 0.5 \cdot Valence$), as well as physical-arousal in CCC for MuSe-Wilder, MuSe-Stress, and MuSe-Physio on the devel(opment) and test partitions. For MuSe-Sent, we report F1 score across five classes (20 % by chance). As feature sets, we test DeepSpectrum, VGGish, and eGeMAPS for audio; Xception, VGGface and FAU for video; and BERT for text. All utilised features are aligned to the label timestamps by imputing missing values or repeating the word embeddings

Features	MuSe-Wilder			MuSe-Sent			MuSe-Stress			MuSe-Physio
	Valence	Arousal	Combined	Valence	Arousal	Combined	Valence	Arousal	Combined	
	devel / test	devel / test	devel / test	devel / test	devel / test	devel / test	devel / test	devel / test	devel / test	devel / test
Audio										
DeepSpectrum	.1901 / .1019	.4841 / **.3386**	.3371 / .2203	30.23 / 27.26	33.52 / 33.16	31.88 / 30.21	.5018 / .4525	.3091 / .2341	.4055 / .3433	.4423 / .4162
VGGish	.1500 / .0054	.4027 / .2545	.2764 / .1300	30.76 / 25.08	36.05 / 31.66	33.41 / 28.37	.5370 / .4766	.1348 / .0296	.3359 / .2531	.3180 / .3967
eGeMAPS	.1916 / .0019	.3877 / .2428	.2897 / .1224	32.93 / 25.80	36.04 / 31.97	34.49 / 28.89	.5845 / .5018	.4304 / .4416	.5075 / .4717	.3381 / .2416
Video										
Xception	.1872 / .1637	.2870 / .1793	.2371 / .1715	30.40 / 28.74	35.16 / 31.14	32.78 / 29.94	– / –	– / –	– / –	– / –
VGGface	.1203 / .1197	.3201 / .2970	.2202 / .2084	32.29 / 28.86	34.57 / 31.32	33.43 / 30.09	.4653 / .4529	.2004 / .1579	.3329 / .3054	.3903 / .4582
FAU	.0682 / .1275	.3045 / .1165	.1864 / .1220	31.37 / 27.38	35.21 / 31.43	33.29 / 29.41	.3565 / .2731	.3313 / .2641	.3439 / .2686	.3344 / .1404
Text										
Bert	.4613 / .5671	.2716 / .1873	.3665 / .3772	32.68 / 31.90	**38.27** / 30.63	**35.48** / 31.27	.2619 / .1747	.2334 / .1446	.2477 / .1597	.2583 / .1604
Late Fusion										
best A + V	.2362 / .1220	.4821 / .2822	.3592 / .2021	**32.96** / 27.92	37.72 / **35.12**	35.34 / 31.52	**.6966 / .5614**	**.5043 / .4562**	**.6005 / .5088**	**.4913 / .4908**
best A + T	.4782 / .5950	.4754 / .3046	.4768 / .4498	30.15 / 30.29	37.63 / 32.87	33.89 / 31.58	.5684 / .5192	.4589 / .3227	.5137 / .4210	.3931 / .1758
best V + T	.4641 / .5874	.3111 / .1767	.3876 / .3821	30.17 / **32.91**	37.51 / 32.73	33.84 / **32.82**	.5588 / .4250	.2891 / .1586	.4240 / .3828	.2734 / .3000
best V + A + T	**.4863 / .5974**	**.4929** / .3257	**.4896 / .4616**	30.37 / 31.01	36.72 / 33.20	33.55 / 32.11	.6769 / .5349	.4819 / .3472	.5794 / .4411	.4330 / .3205

Using the one-dimensional biological signals as features might also be beneficial here, even though our model fails to generalise for them. We achieve CCCs of .4188 on the development and .3328 on the test set using a 4 LSTM layer setting and a learning rate of 0.01.

5 CONCLUSIONS

In this paper, we introduced MuSe 2021 – the second Multimodal Sentiment Analysis challenge. MuSe 2021 utilises the MuSe-CaR multimodal corpus of emotional car reviews and the Ulm-TSST corpus, including bio-signals, which are newly featured for the MuSe challenge. The 2021 challenge is comprised of four sub-challenges, aimed for predicting in: i) MuSe-Wilder, the level of the affective dimensions of valence (corresponding to sentiment) and arousal; ii) MuSe-Sent, five classes of each, valence and arousal, from video parts containing certain topics; iii) MuSe-Stress, the level of continuous valence and arousal from stressful situations; and iv) MuSe-Physio a combination of arousal and EDA signals. By intention, we decided to use open-source software to extract a wide range of feature sets to deliver the highest possible transparency and realism for the baselines. Besides the features, we also share the raw data and the developed code for our baselines publicly. The official baseline for each sub-challenge is for MuSe-Wilder .5974 for continuous valence using late fusion and .3386 for continuous arousal using DeepSpectrum features; for the five-class classification MuSe-Sent, an F1 score of 32.91 % for valence utilising late fusion of vision and text and 35.12 % for arousal utilising a late fusion of audio-video; for MuSe-Stress, a CCC of .5614 for valence and .4562 for arousal, both based on fusion of the best audio and visual features; and finally, for MuSe-Physio, a CCC of .4908 for physiological-emotion prediction.

The baselines are improved through the use of a simple fusion method and show the challenge ahead for multimodal sentiment analysis. In the participants' and future efforts, we hope for novel and exciting combinations of the modalities – such as linking modalities at earlier stages in the pipeline or more closely.

6 ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS). We thank the sponsors of the Challenge, the BMW Group, and audEERING.

REFERENCES

[1] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn W Schuller. 2017. Snore Sound Classification Using Image-Based Deep Spectrum Features.. In *Proceedings of INTERSPEECH*, Vol. 434. 3512–3516.

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2020. Gated multimodal networks. *Neural Computing and Applications* (2020), 1–20.

[3] 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 (MMSP)*. IEEE, 1–5.

[4] 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, Changu, China.

[5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: an Open Source Facial Behavior Analysis Toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE.

[6] Yekta Said Can, Bert Arnrich, and Cem Ersoy. 2019. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of Biomedical Informatics* 92 (2019).

[7] Delphine Caruelle, Anders Gustafsson, Poja Shams, and Line Lervik-Olsen. 2019. The use of electrodermal activity (EDA) measurement to understand consumer emotions–a literature review and a call for action. *Journal of Business Research* 104 (2019), 146–160.

[8] 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*. 4171–4186.

[9] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. 2020. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI)*. 784–789.

[10] Richard O Duda, Peter E Hart, et al. 1973. *Pattern classification and scene analysis.* Vol. 3. Wiley New York.

[11] J. O. Egede, S. Song, T. A. Olugbade, C. Wang, A. C. D. C. Williams, H. Meng, M. Aung, N. D. Lane, M. Valstar, and N. Bianchi-Berthouze. 2020. EMOPAIN Challenge 2020: Multimodal Pain Evaluation from Facial and Bodily Expressions. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020).* 849–856.

[12] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202.

[13] 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.* 1459–1462.

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 776–780.

[15] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *The IEEE Winter Conference on Applications of Computer Vision.* 1470–1478.

[16] Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 381–385.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

[18] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 131–135.

[19] Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht, Harald C Traue, and Henrik Kessler. 2012. Mapping discrete emotions into the dimensional space: An empirical approach. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC).* IEEE, 3316–3320.

[20] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. 1993. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.

[21] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. 2020. Analysing affective behavior in the first ABAW 2020 competition. *arXiv preprint arXiv:2001.11409* (2020).

[22] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. 2010. Understanding of internal clustering validation measures. In *Proceedings of the IEEE International Conference on Data Mining.* IEEE, 911–916.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV).*

[24] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (1982), 129–137.

[25] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.. In *Proceedings of INTERSPEECH*, Vol. 2017. 498–502.

[26] Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion Measurement.* Elsevier, 201–237.

[27] Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic. 2014. Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1299–1311.

[28] Yannis Panagakis, Mihalis A Nicolaou, Stefanos Zafeiriou, and Maja Pantic. 2015. Robust correlated and individual component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2015), 1665–1678.

[29] Vedhas Pandit and Björn Schuller. 2019. On Many-to-Many Mapping Between Concordance Correlation Coefficient and Mean Square Error. *arXiv preprint arXiv:1902.05180* (2019).

[30] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference (BMVC).* 41.1–41.12.

[31] Sara Pourmohammadi and Ali Maleki. 2020. Stress detection using ECG and EMG signals: A comprehensive study. *Computer Methods and Programs in Biomedicine* 193 (2020), 105482.

[32] Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.* 9–15.

[33] Xiaoyu Qiu, Zhiquan Feng, Xiaohui Yang, and Jinglan Tian. 2020. Multimodal Fusion of Speech and Gesture Recognition based on Deep Learning. In *Journal of Physics: Conference Series*, Vol. 1453.

[34] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. 2018. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop.* 3–13.

[35] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge.* 3–9.

[36] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.

[37] Björn W Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, et al. 2021. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, escalation & primates. *arXiv preprint arXiv:2102.13468* (2021).

[38] Björn W Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, et al. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. *Proceedings of INTERSPEECH* (2020).

[39] Björn W Schuller, Stefan Steidl, Anton Batliner, Peter B Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian B Pokorny, Eva-Maria Rathner, Katrin D Bartl-Pokorny, et al. 2018. The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats.. In *Proceedings of INTERSPEECH.* 122–126.

[40] Jainendra Shukla, Miguel Barreda-Angeles, Joan Oliver, GC Nandi, and Domenec Puig. 2019. Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing* (2019).

[41] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[42] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjoern W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-Target Engagement and Trustworthiness Detection in Real-Life Media. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop.* ACM, 35–44.

[43] Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. 2021. The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements. *IEEE Transactions on Affective Computing (Early Access)* (June 2021). https://doi.org/10.1109/TAFFC.2021.3097002

[44] Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL. *arXiv preprint arXiv:2004.13850* (2020).

[45] Lukas Stappen, Vincent Karas, Nicholas Cummins, Fabien Ringeval, Klaus Scherer, and Björn Schuller. 2019. From speech to facial activity: towards cross-modal sequence-to-sequence attention networks. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP).* IEEE, 1–6.

[46] Lukas Stappen, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. 2021. MuSe 2021 Challenge: Multimodal Emotion, Sentiment,Physiological-Emotion, and Stress Detection. In *29th ACM International Conference on Multimedia (ACMMM).* ACM, Virtual Event, China.

[47] Lukas Stappen, Georgios Rizos, Madina Hasan, Thomas Hain, and Björn W Schuller. 2020. Uncertainty-Aware Machine Support for Paper Reviewing on the INTERSPEECH 2019 Submission Corpus. *Proceedings of INTERSPEECH* (2020), 1808–1812.

[48] Lukas Stappen, Björn Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. Summary of MuSe 2020: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media. In *Proceedings of the 28th ACM International Conference on Multimedia.* 4769–4770.

[49] Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigel, Erik Cambria, and Björn W Schuller. 2021. MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge, co-located with the 29th ACM International Conference on Multimedia (ACMMM).* ACM, Changu, China.

[50] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. 2020. Multimodal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop.* 27–34.

[51] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2544–2558.

[52] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–10.

[53] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. 2008. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of INTERSPEECH*. 597–600.

[54] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2015. WIDER FACE: A Face Detection Benchmark. *CoRR* abs/1511.06523 (2015). arXiv:1511.06523 http://arxiv.org/abs/1511.06523

[55] Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, and Soujanya Poria (Eds.). 2020. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*.

[56] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23 (04 2016).

[57] Feng Zhou and Fernando De la Torre. 2015. Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 2 (2015), 279–294.