Comparing Biosignal and Acoustic feature Representation for Continuous Emotion Recognition

Sarthak Yadav
Idiap Research Institute
Martigny, Switzerland
sarthak.yadav@idiap.ch

Tilak Purohit
Idiap Research Institute, Martigny
École polytechnique fédérale de Lausanne (EPFL)
Switzerland

Zohreh Mostaani
Idiap Research Institute, Martigny
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

Bogdan Vlasenko
Idiap Research Institute
Martigny, Switzerland

Mathew Magimai.-Doss
Idiap Research Institute
Martigny, Switzerland

ABSTRACT

Automatic recognition of human emotion has a wide range of applications. Human emotions can be identified across different modalities, such as biosignal, speech, text, and mimics. This paper is focusing on time-continuous prediction of level of valence and psycho-physiological arousal. In that regard, we investigate, (a) the use of different feature embeddings obtained from neural networks pre-trained on different speech tasks (e.g., phone classification, speech emotion recognition) and self-supervised neural networks, (b) estimation of arousal and valence from physiological signals in an end-to-end manner and (c) combining different neural embeddings. Our investigations on the MuSe–Stress sub-challenge shows that (a) the embeddings extracted from physiological signals using CNNs trained in an end-to-end manner improves over the baseline approach of modeling physiological signals, (b) neural embeddings obtained from phone classification neural network and speech emotion recognition neural network trained on auxiliary language data sets yield improvement over baseline systems purely trained on the target data, and (c) task specific neural embeddings yield improved performance over self-supervised neural embeddings for both arousal and valence. Our best performing system on test-set surpass the DeepSpectrum baseline (combined score) by a relative 7.7% margin.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

KEYWORDS

Emotion recognition, Self-supervised embedding, pre-trained embedding, modalities fusion, breathing patterns

MuSe' 22, October 10, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9484-0/22/10…$15.00
https://doi.org/10.1145/3551876.3554812

ACM Reference Format:
Sarthak Yadav, Tilak Purohit, Zohreh Mostaani, Bogdan Vlasenko, and Mathew Magimai.-Doss. 2022. Comparing Biosignal and Acoustic feature Representation for Continuous Emotion Recognition. In *Proceedings of the 3rd International Multimodal Sentiment Analysis Workshop and Challenge (MuSe' 22), October 10, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3551876.3554812

1 INTRODUCTION

Emotions are quintessential elements of communication among humans, and are expressed in different ways across several modalities. Speech is one of the prime modes to convey the expression of emotions, and hence emotion recognition using the acoustic content of signal is gaining popularity in speech application areas. Human emotions are paralinguistic phenomena which manifest distinctively over varying temporal and spectral characteristics. Due to limitations with representation and processing, extraction of human emotions using traditional acoustic signal analysis method is challenging. Recent trends have witnessed a growing interest in the field of multimodal emotion recognition, which attracts extensive use of deep neural networks (DNNs) to exploit the contrast between speech, textual, and physiological modalities [1, 2, 3, 4, 5, 6, 7].

Physiological signals have been used for emotion recognition [8]. Within the field of affective computing, recognition approaches to predict continuous states of emotion, frequently utilize the two-dimensional Circumplex Emotion Model [9], observing the valence and arousal of speaker emotional state. However, in order to avoid subjective emotional labels, multiple raters must continuously annotate, which is costly and time-expensive. In [10] authors showed that bio-signals processed with MuSe-Toolbox could be used as alternative to Evaluator Weighted Estimator (EWE) [11] emotional gold standard [12]. Experimental study presented by [13] shows that respiration patterns reflect the general dimensions of emotional response.

The '*Multimodal Sentiment Analysis in Real-life Media*' (MuSe) 2022 challenge provides an opportunity for researchers to evaluate emotion recognition systems using multiple modalities (audio, biosignals, video, and text) over in–the–wild data. This paper focuses on the MuSe–Stress task sub–challenge, where the goal is to predict emotion in a time continuous manner. In that context, we investigate,

(1) Modeling of different embeddings extracted from supervised neural networks, i.e. trained on specific speech tasks on auxiliary data such as, phone recognition, speech emotion recognition, speech breathing pattern estimation and from neural networks trained in self-supervised manner on auxiliary data.
(2) Modeling of physiological signals in an end-to-end manner using convolutional neural networks.
(3) Fusion of different feature embedding spaces.

We demonstrate the potential of these approaches by contrasting them against the baseline systems developed by the challenge organizers.

The rest of the paper is organized as follows. Section 2 describes the feature representations derived from the acoustic and biosignal modalities, utilized in the subsequent studies. Section 3 describes the database, evaluation metrics, methods utilized to create the proposed systems, and the experimental setup of the proposed systems. Section 4 presents the results obtained using different methods and systems. Section 5 concludes the paper.

2 PROPOSED APPROACHES

We pursued two approaches, namely, (a) modeling embeddings extracted from acoustic signal using pre–trained neural networks (Figure 1) and (b) an end-to-end CNN-based system modeling physiological signals (Figure 2) to estimate valence and arousal. Furthermore, we investigated fusion of acoustic information and physiological information.

2.1 Pre-trained feature representations

Figure 1 illustrates our method. In this approach, first, frame-level neural embeddings are extracted from pre–trained networks using the acoustic signal. Fixed length representation is then obtained for each 500 ms of signal by applying functionals namely *mean* and *standard deviation (std)*. Arousal and valence are then estimated by feeding the fixed length representation as input to a neural network.

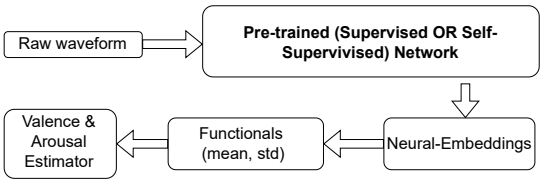

Figure 1: Proposed pipeline for using embeddings from pre-trained networks.

In the remainder of the section, we first present the different embeddings investigated. They can be broadly grouped as supervised and self-supervised neural network embeddings. The supervised neural network embeddings are extracted by training neural networks in a task specific manner. We then present the arousal and valence estimator.

2.1.1 Phonetic feature representations (denoted as Raw(PHN)). Phonetic information has been shown to capture emotional content in speech. Previously, [14] showed a strong correlation between the vowel formants and the level of arousal in human speech, while [15] demonstrated that incorporating speech-articulatory information improves valence-based classification. The benefit of modelling speech units for speech emotion recognition (SER) task was shown in [16]. Recently, [17] exhibited through their work that phonetic units are beneficial and should be incorporated for SER based studies. All these prior works inspired us to use phone-based embedding for the task-on-hand. For this we make use of an off-the-shelf CNN based network that models raw-audio signals for phoneme-classification. The network was trained on the AMI Meeting corpus [18] containing 100h of meetings. The network takes 250ms of raw audio with a 10ms shift as an input, and consists of 10 convolutional layers followed by a fully-connected layer with 1024 neurons and an output unit. The model provides neural embeddings of dimension 1024 corresponding to each 250ms frame, these embeddings are then converted to fixed-length utterance representations by computing functionals. The resulting representation denoted as, Raw(PHN) is 2048 (1024 for mean and 1024 for std) dimensional vector.

2.1.2 Speech-based emotion recognition feature representation (denoted as Raw(SER)). Since the task-on-hand deals with predicting two of the emotion dimensions, emotional-valence and arousal, it deemed appropriate to generate embeddings from a network trained for SER task. For this we resort to an off-the-shelf CNN network similar to the Raw-waveform CNN network presented in [19] that models raw-audio signals. The network was trained for SER task using IEMOCAP corpus [20], a benchmark database for SER. The networks was trained in a end-to-end manner, where input to the system was 250ms of raw audio signal, the network consists of four convolutional layers followed by a fully connected layer with ten nodes and an output layer with softmax activation for a four class classification corresponding to four emotion categories namely- sad, happy, angry, and neutral. The network was trained in a speaker-independent fashion. The frame level neural embeddings of dimension 10 were extracted using this network. The fixed-length representation after applying functionals, denoted as Raw(SER) is a 20 (10 for mean and 10 for std) dimensional vector.

2.1.3 Breathing pattern embedding (denoted as UCLBS). Speech carries a wide range of information including age [21], gender [22], and emotional state [23] of a person. Respiration is one of the physiological signals altered by emotion. Relationships between emotions and respiratory patterns have shown more rapid breathing during an speaker's emotional arousal state [24]. Respiration with deep learning based methods has been used for emotion recognition [25]. There is a close relation between speech and breathing as well since speech is produced by organs evolved for breathing [26]. Recently, estimating breathing patterns from speech signals has gained more attention. [27] used log Mel Spectrogram of speech to train a CNN and a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to predict the breathing signal [27]. The Interspeech 2020 ComParE challenge, was partly dedicated to estimating breathing patterns from speech signals and several methods including end-to-end systems were proposed [28, 29]. Estimating breathing patterns from raw speech waveform using CNN was studied in [30, 31]. It has been shown that the embeddings extracted from such CNNs pre–trained for estimating breathing pattern are informative for

auxiliary tasks such as detection of COVID–19 [32] and distinguishing between natural and synthetic speech [33]. Based on these observations we hypothesise that such embeddings can be used for emotion recognition.

For this purpose we used an off-the-shelf CNN trained for estimating breathing patterns in the output by taking 3 seconds of speech signals in the input. The CNN consists of 4 convolution layers followed by a fully connected layer with 10 nodes and finally an output layer. The network was trained using UCL-SBM database [28] consists of conversational speech and Mean Squared Error was used as loss function during training. More details about the pre-trained networks can be found in [30].

The embeddings are extracted before the activation of the fully connected layer for every 20ms. The fixed-length representation after applying functionals, denoted as UCLBS is a 20 (10 for mean and 10 for std) dimensional vector.

2.1.4 Self-supervised learning (SSL) representations (denoted as COLA, HuBERT and WavLM). Several recent works propose learning general purpose acoustic representations in a self-supervised fashion [34, 35, 36, 37, 38, 39, 40, 41, 42, 43]. While these methods learn representations that yield excellent few-shot and linear probe performance on several diverse downstream audio classification tasks, the viability of these representations for continuous emotional recognition is yet to be evaluated. In our experimental study we utilize two types of SSL embeddings: general purpose audio and so-called "full stack" speech processing representations.

The general purpose representations are trained on AudioSet [44] using the COLA [40] framework for contrastive self-supervised learning. AudioSet is the largest publicly available audio event dataset, with over 2M 10-sec clips spanning over 632 audio event classes. COLA learns a latent space where the similarity between anchor-positive pair of audio segments from the same audio clip is greater than the similarity between anchor segment and other negative distractors by optimizing the following objective function:

$$\mathcal{L} = -\log \frac{\exp\left(\mathrm{s}(x, x^{+})\right)}{\sum_{x^{-} \in X^{-}(x) \cup \{x^{+}\}} \exp\left(\mathrm{s}(x, x^{-})\right)}, \tag{1}$$

where $s(x, x') = g(f(x))^{\top}\ W\ g(f(x'))$ is the similarity function, (x, x^{+}) denotes the anchor-positive pair, $X^{-}(x)$ refers to the set of negative samples, $f(.)$ represents the lightweight EfficientNet-B0 [45] convolutional feature encoder, $g(.)$ is a shallow neural network that maps input features h onto a space $z = g(h) \in \mathbb{R}^{G}$, and $W \in \mathbb{R}^{G}$ are bilinear similarity parameters. In line with previous experimental protocol [40, 43], we use the 1280 dimensional feature embedding returned by the EfficientNet-B0 encoder trained on Mel-spectrogram features for the provided audio signals for our experiments, which is available publically [46], and is here on referred to simply as COLA.

Finally, the HuBERT [47] and WavLM [48] were served as the "full-stack" speech processing representations. Systems build on top of these embedding are among the top three performing networks for the SUPERB challenge [49], an SSL benchmark challenge for the speech processing tasks. WavLM Large, which was pre-trained on 94k hours of speech, and HuBERT Large, which was pre-trained on 60k Libri-light speech, were used in the presented experimental studies.

2.1.5 Arousal and valence estimator. The sequential nature of the selected regression tasks makes recurrent neural networks (RNNs) a natural choice for a comparably simple system. We used the LSTM-RNN system proposed as part of baseline system without any modification to the architecture or training process for estimating valence and arousal. As done in the baseline studies, we train a separate estimator for valence and arousal. It allows us to fairly compare our proposed embeddings and feature representations to those of the baseline studies.

2.2 Modelling raw physiological signals using CNNs

Several works propose modelling raw waveform signals for various tasks, ranging from speech recognition [50, 51, 52], speaker recognition [53, 54], gender recognition [55], depression detection [56], audio classification [57], as well as for modelling raw physiological signals [30, 31, 33], among others. In the same light, we propose a CNN based framework for directly modelling raw Respiratory, ECG and BPM physiological signals for estimating valence and arousal in an end-to-end manner. Figure 2 illustrates this method. Given the nature of these input signals as well as the annotation granularity of the data (every 500 ms), each physiological signal is modelled after centering of the labelled segment with appropriate context.

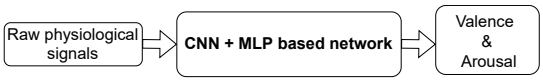

Figure 2: Proposed pipeline for end-to-end system.

Table 1 depicts the general CNN architecture used for modelling respiratory (Resp-CNN), BPM (BPM-CNN) and ECG (ECG-CNN) physiological signals. The proposed CNN architecture comprises of 4 convolutional layers followed by an MLP with one hidden layer. All the hidden layers are followed by a ReLU activation function. The number of filters in each layer as well as the kernel and stride parameters of the first convolution layer, which are dependent on the signal characteristics, were tuned individually for each physiological signal. A Dropout layer [58] was added before the MLP for improved regularisation. More information on the ablation experiments on the development set is presented in Section 3.4.

2.3 Fusion-based estimation

We also investigate a combination of different embeddings extracted from acoustic signal and physiological signals. We investigate early fusion, where different features are concatenated and are fed as input to the RNN-based arousal and valence estimator, presented earlier in Section 2.1.5.

3 EXPERIMENTAL SETUP

The section describes the experimental setup for the proposed study, including a brief description of the dataset and the evaluation protocol used and the evaluation metric used for the study. This is

Table 1: CNN architectures for physiological signals. Conv parameters are denoted as (filters, kernel width, stride), and MP denotes MaxPooling layer

Resp-CNN	BPM-CNN	ECG-CNN
Conv(64, 75, 15)	Conv(16, 175, 10)	Conv(56, 100, 15)
MP(2, 2)	MP(2, 2)	
Conv(128, 10, 1)	Conv(32, 10, 1)	Conv(112, 10, 1)
MP(2, 2)	MP(2, 2)	
Conv(256,7,1)	Conv(64,7,1)	Conv(224,7,1)
Conv(512,7,1)	Conv(128,7,1)	Conv(448,7,1)
FC(75)	FC(175)	FC(100)

followed by an in-depth description of the training methodology and ablation experiments for development of the proposed novel raw physiological signal modelling CNNs.

3.1 Dataset and protocol

The MuSe–Stress sub–challenge is a regression task on continuous signals for emotional arousal and valence [59]. The Ulm-Trier Social Stress Test dataset (Ulm-TSST) is used to setup training, development and test sub-sets, comprising individuals in a stressful situations following the Trier Social Stress Test (TSST) [60]. The dataset provides *training*, *development* and *testing* splits with 41, 14 and 14 subjects, respectively. We further split the training set into *train* set with 32 subjects and *validation* set with 9 subjects for tuning hyper-parameters.

3.2 Evaluation metrics

The regression task MuSe–Stress is evaluated in terms of Concordance Correlation Coefficients (CCC) for arousal, valence and combined modalities. The ultimate goal of the challenge is to reach the highest possible combined *CCC* score.

3.3 Baseline systems

The challenge organizers provide systems which can be used to evaluate the classification performance of features and networks for emotion dimensions. Also, extracted feature representation for audio (DeepSpectrum [61]) and bio-signal (BPM, ECG and respiratory signal) were provided by organizers.

3.4 Training raw physiological signal CNNs

In this section we describe the process of training the proposed physiological signal CNNs. For the challenge, the provided raw physiological signals have a sampling frequency of 1000 Hz, and labels are provided for every 500 ms intervals. Each physiological signal is centered with a *context* such that the center 500 ms segment of the input signal corresponds to the target label, normalized and is then directly fed to the model. We adopt a multi-task learning paradigm for optimizing physio-arousal and valence simultaneously, i.e., the final layer of each CNN returns 2 outputs, and the CNN is trained by optimizing the average combined CCC. Each CNN is trained in a multi-task setting with an AdamW optimizer [62], with early stopping.

Table 2: Context-wise ablation study for physiological CNNs. Combined CCC score on the development set is reported.

Model	Context (ms)	Combined [CCC] Development
Resp-CNN	0	0.1259 (0.1144±0.006)
	500	0.1389 (0.1256±0.0145)
	1000	0.0938 (0.0813±0.0089)
	1500	0.1432 (0.1167±0.0124)
	2000	0.1504 (0.1271±0.0112)
	2500	**0.1764 (0.1509±0.0160)**
	3000	0.1527 (0.1266±0.0202)
ECG-CNN	0	0.2067 (0.1890±0.0120)
	250	0.3657 (0.3170±0.0294)
	500	**0.3885 (0.3280±0.0296)**
	1000	0.3514 (0.2908±0.0357)
	1500	0.3447 (0.3064±0.0253)
BPM-CNN	0	0.114 (0.0913±0.011)
	500	0.137 (0.106±0.0233)
	1000	**0.2224 (0.2042±0.0091)**

Given the different characteristics of each physiological signal for each physiological signal CNN, we optimized the following hyperparameters:

(1) *context*: We ran individual ablation experiments to yield the context that gives the best performance.
(2) *Number of parameters*: We separately optimized complexity of the physiological CNNs by tuning a *multiplier* hyperparameter that scales the number of filters in each layer.
(3) *FC layer dimensions*: Dimensions of the first fully connected layer is individually tuned for each physiological signal CNN.

Table 2 shows the results of the context-wise ablation study for training raw physiological system CNNs. Respiratory signal works best with the largest context of 2500 ms, which is inline with observations made in recent literature for breathing pattern estimation where similarly large context sizes yield better performance [31, 33], whereas apt contexts for other settings are inline with characteristics of the underlying physiological signal.

4 RESULTS AND ANALYSIS

This section describes the results obtained using various uni- and multi-modal feature representations. Table 3 shows the results of the various systems on the MuSe–Stress development and test set. The top rows depict the best baseline result (DeepSpectrum) and the best physiological signal based system as per the challenge white paper [59], followed by our results using features obtained from the proposed physiological CNNs and acoustic representations. Finally, select multi-modal early-fusion results are provided. All methods described are built on top of extracted features and their combinations using the provided baseline code.

4.1 Uni–modal systems

4.1.1 Modelling acoustic signals. From Table 3, it could be observed that the proposed acoustic embeddings, Raw(SER) and Raw(PHN)

Table 3: *CCC* **scores obtained on the development and the test set by various systems. Best scores over 5 random seeds reported, with (mean ± std) over runs for the development set. "+" denotes early fusion, i.e. concatenation of the denoted features, respectively.** ***ndims*** **denotes feature dimensionality.**

		Arousal [CCC]		Valence [CCC]		Combined [CCC]
Features	ndims	Development	Test	Development	Test	Test
Baseline systems						
DeepSpectrum	1024	0.4139 (0.3433 ± 0.0548)	**0.4239**	0.5741 (0.5395 ± 0.0207)	**0.4931**	**0.4585**
eGeMAPS	88	0.4112 (0.3168 ± 0.0459)	0.2975	0.5090 (0.4744 ± 0.0244)	0.3988	0.3482
BPM + ECG + Resp	3	0.3917 (0.2793 ± 0.0782)	0.1095	0.4361 (0.2906 ± 0.0787)	0.1861	0.1478
Proposed systems						
Physiological						
UCLBS	20	0.1606 (0.1356 ± 0.0149)	0.0794	0.3994 (0.3286 ± 0.0410)	**0.3044**	0.1920
Resp-CNN+ECG-CNN+BPM-CNN	350	0.4315 (0.3899 ± 0.0442)	0.1340	0.5445 (0.5323 ± 0.0130)	0.1814	0.1577
UCLBS+ECG-CNN+BPM-CNN	295	0.4333 (0.3749 ± 0.0421)	**0.1890**	0.5794 (0.5505 ± 0.0219)	0.2595	**0.2242**
Acoustic						
Raw(SER)	20	0.3404 (0.2986 ± 0.0311)	0.4338	0.5548 (0.5403 ± 0.0116)	**0.5134**	0.4736
Raw(PHN)	2048	0.3515 (0.3371 ± 0.0102)	0.4909	0.4122 (0.3894 ± 0.0217)	0.4767	0.4838
COLA	1280	0.3770 (0.3480 ± 0.0266)	0.4764	0.5572 (0.5268 ± 0.0310)	0.3028	0.3896
HuBERT	2048	0.2622 (0.2388 ± 0.0155)	0.4833	0.5098 (0.4853 ± 0.0161)	0.4309	0.4571
WavLM	2048	0.2842 (0.2599 ± 0.0183)	0.4462	0.4672 (0.4381 ± 0.0240)	0.4874	0.4668
Raw(SER)+Raw(PHN)	2068	0.3742 (0.3540 ± 0.0176)	0.4850	0.4081 (0.3804 ± 0.0214)	0.4966	0.4908
Raw(SER)+COLA	1300	0.3818 (0.3593 ± 0.0241)	**0.5111**	0.5528 (0.5429 ± 0.0082)	0.4023	0.4567
Raw(PHN)+COLA	3328	0.3860 (0.3558 ± 0.0356)	0.4014	0.4335 (0.4081 ± 0.0204)	0.4822	0.4418
Raw(SER)+HuBERT	2068	0.3144 (0.3063 ± 0.0063)	0.4724	0.4941 (0.4630 ± 0.0185)	0.4907	0.4815
Raw(SER)+WavLM	2068	0.3114 (0.2924 ± 0.0134)	0.4354	0.4587 (0.4500 ± 0.0065)	0.4648	0.4501
Raw(PHN)+HuBERT	4096	0.3620 (0.3466 ± 0.0107)	0.4743	0.4395 (0.4098 ± 0.0183)	0.4713	0.4728
Raw(PHN)+WavLM	4096	0.3736 (0.3614 ± 0.0114)	0.4590	0.4375 (0.4114 ± 0.0174)	0.4525	0.4557
Raw(SER)+Raw(PHN)+HuBERT	4116	0.3683 (0.3441 ± 0.0187)	0.4749	0.4191 (0.4019 ± 0.0179)	0.4316	0.4533
Raw(SER)+Raw(PHN)+WavLM	4116	0.3670 (0.3421 ± 0.0213)	0.4909	0.4237 (0.4083 ± 0.0117)	0.4304	0.4607
Raw(SER)+Raw(PHN)+COLA	3348	0.3697 (0.3618 ± 0.0045)	0.4767	0.4356 (0.4208 ± 0.0151)	0.5109	**0.4938**
Multi-modal early fusion						
UCLBS+Raw(SER)	40	0.4382 (0.3700 ± 0.0506)	0.3218	0.5602 (0.5273 ± 0.0222)	0.3597	0.3407
UCLBS+Raw(PHN)	2068	0.3803 (0.3579 ± 0.0189)	0.4644	0.4529 (0.4027 ± 0.0258)	**0.4952**	**0.4798**
Raw(SER)+Raw(PHN)+DeepSpectrum	3092	0.3764 (0.3490 ± 0.0257)	**0.4734**	0.4280 (0.4114 ± 0.0195)	0.4386	0.4560

are able to surpass the best performing DeepSpectrum baseline results on the test set for both valence and arousal, also the overall result obtained via these embeddings on the test set are outperforming the best performing baseline systems. The hypothesis that the task-on-hand deals with speech emotion and SER based embeddings might help seems correct, furthermore it is interesting to observe that although our SER system was trained on IEMOCAP [20], an English corpora, the embeddings derived from it (Raw(SER)) generalised well for MuSe–Stress data which is recorded in German language. This highlights the potential of our training methodology of modelling sub-segmental speech (typically 250ms), which made the network and therefore its derived embeddings robust towards new language. It is also worth noting that the Raw(SER) embeddings despite only being 20 dimensional provide the best standalone results for emotional-valence prediction. Similar to Raw(SER), Raw(PHN) embeddings also appears to be robust towards unseen language, given that it generalises well for the MuSe–Stress data despite the network for deriving Raw(PHN) being trained on an English language corpora. These results also showcase that phonetic level information is crucial and complement emotion recognition task, in line with the observations made by previous studies mentioned in subsection 2.1.1. It is worth mentioning that phonetic embeddings also showed good results for the case of non-speech vocalizations [63] outperforming the DeepSpectrum baseline for the ExVo multi-learning task [64]. The Raw(PHN) embedding gives the best results for the emotional-arousal prediction for a standalone embedding. Moreover, it is interesting to see that the Raw(PHN) and Raw(SER) embeddings complement one another, with early fusion results of these embeddings providing a superior overall score.

In our experiments, standalone HuBERT and WavLM features also perform quite well, with the latter outperforming the baseline DeepSpectrum result. However, when used in an early fusion setting with other feature sets, we observe a reduction in combined test performance, which is more pronounced for the WavLM features.

Finally, the Raw(SER)+Raw(PHN)+COLA system reaches a combined test performance of 0.4938, which is an approx. 8% relative

improvement over the best DeepSpectrum baseline result, and is our best performing system. It is worth mentioning that this performance is achieved using COLA while COLA is the worse performing standalone acoustic embedding, demonstrating that they are synergistic with phonetic and emotion embeddings for the task-at-hand.

4.1.2 Modelling physiological signals. From Table 3, we can see that all of our proposed physiological modelling methods outperform the physiological baseline from [59]. The fusion of the features extracted from the proposed physiological CNNs (Resp-CNN+ECG-CNN+BPM-CNN) improves test performance over the baseline, signifying the viability of directly modelling raw physiological signals.

It's worth noting that the feature embeddings extracted from the breathing pattern estimation model (UCLBS), while performing slightly worse for arousal estimation in comparison to the baseline (0.0794 v/s 0.1095), significantly outperforms both the baseline and the proposed physiological CNNs for valence as well as the combined test CCC performance. A possible explanation for this phenomena is the fact that these embeddings are extracted from a pre-trained network trained on raw waveforms from a conversational speech database, and thus potentially include speech-related discriminative information which have been demonstrated to be informative for other tasks [32, 33], further boosting their viability. We also note that, similar to the systems used for modeling acoustic features this network is also pre–trained on English database and is generalizing well on Muse–Stress data which is in German language.

Given the better performance of UCLBS features, we decided to replace the Resp-CNN embeddings and training a fusion of UCLBS features with the other physiological CNNs (*UCLBS+ECG-CNN+BPM-CNN*), which, by trading off the excellent valence performance of UCLBS features for a significant improvement in arousal performance results in a 50% relative improvement in combined test score over the physiological-only baseline (0.2242 v/s 0.1478). However, it's worth noting that there is still a very large discrepancy between development and test performance for physiological only systems, highlighting that these systems struggle with overfitting on training distribution.

4.2 Multi–modal systems

Following results of uni–modal systems, we experiment with select multi–modal early fusion approaches of the top performing systems across modalities. First, we fused UCLBS breathing estimation features, which were the best performing standalone physiological feature set, with Raw(PHN) and Raw(SER) features, which are our top performing uni–modal features. However, the subsequent fused features showed performance degradation over the constituent acoustic features, with a much larger degradation observed for Raw(PHN). We also fuse Raw(SER) and Raw(PHN) with the best baseline feature representation (DeepSpectrum), which also does not improve the performance.

5 CONCLUSIONS

For MuSe 2022 stress sub-challenge, we investigated modeling of different feature embeddings obtained from task-specific pre-trained neural networks and self-supervised networks, as well as modeling of physiological signals for the continuous estimation of arousal and valence. Multi modal systems were investigated by using early fusion between different modalities. Additionally we investigated modeling of physiological signals in an end-to-end manner for the task-at-hand. While the proposed physiological models outperform the physiological baseline [59], embeddings extracted from pre–trained networks perform much better for valence and arousal estimation, including the pre–trained breathing pattern embeddings that model speech signals, demonstrating that acoustic features tend to be more informative for valence and arousal estimation when compared to physiological related information. From Table 3 it could be observed that, among acoustic features, embeddings derived from task-specific pre-trained networks (Raw(SER) and Raw(PHN)) perform better than the self-supervised network based embeddings for the task-at-hand.

Finally, our best performing system is an early fusion of COLA embeddings with pre-trained supervised embeddings (Raw(PHN) + Raw(SER)), which enhances valence estimation performance while maintaining a good estimate for arousal estimation. This increases our best combined score from 0.4736 to 0.4938, surpassing the best performing DeepSpectrum baseline on the test set by a relative 7.7% margin, suggesting that self-supervised COLA embeddings are complementary to pre-trained supervised embeddings. It is also noteworthy that pre-trained supervised networks (viz. Raw(SER) & UCLBS), despite being trained on an English corpora, generalize well for the MuSe-Stress corpus which is in German language. This needs further investigation and would be a part of our future study.

6 ACKNOWLEDGEMENTS

This work was partially funded by the Swiss National Science Foundation through the Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. $40B2 - 0\_194794/1$) and TIPS: Towards Integrated processing of Physiological and Speech signals (grant no. 200021_188754).

REFERENCES

[1] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Bjőrn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11, 8, 1301–1309.

[2] Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, and Bjőrn W Schuller. 2021. Speech emotion recognition using semantic information. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6279–6283.

[3] Bjőrn W Schuller. 2018. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61, 5, 90–99.

[4] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjőrn W Schuller, Iulia Lefter, et al. 2020. Muse 2020 challenge and workshop: multimodal sentiment analysis, emotion-target engagement and trustworthiness detection

in real-life media: emotional car reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, 35–44.

[5] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review. *Information Fusion*, 59, 103–126.

[6] Guang Shen, Riwei Lai, Rui Chen, Yu Zhang, Kejia Zhang, Qilong Han, and Hongtao Song. 2020. WISE: Word-Level Interaction-Based Multimodal Fusion for Speech Emotion Recognition. In *Proc. Interspeech 2020*, 369–373. DOI: 10.21437/Interspeech.2020-3131.

[7] Aparna Khare, Srinivas Parthasarathy, and Shiva Sundaram. 2020. Multi-Modal Embeddings Using Multi-Task Learning for Emotion Recognition. In *Proc. Interspeech 2020*, 384–388. DOI: 10.21437/Interspeech.2020-1827.

[8] 2020. *Emotiw 2020: driver gaze, group emotion, student engagement and physiological signal based challenges. Proceedings of the 2020 International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 784–789. ISBN: 9781450375818. https://doi.org/10.1145/3382507.3417973.

[9] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39, 6, 1161.

[10] Alice Baird, Lukas Stappen, Lukas Christ, Lea Schumann, Eva-Maria Meßner, and Björn W Schuller. 2021. A physiologically-adapted gold standard for arousal during stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 69–73.

[11] 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.

[12] Alice Baird, Shahin Amiriparian, Miriam Berschneider, Maximilian Schmitt, and Björn Schuller. 2019. Predicting biological signals from speech: introducing a novel multimodal dataset and results. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–5.

[13] Frans A. Boiten, Nico H. Frijda, and Cornelis J.E. Wientjes. 1994. Emotions and respiratory patterns: review and critical analysis. *International Journal of Psychophysiology*, 17, 2, 103–128. ISSN: 0167-8760. DOI: https://doi.org/10.1016/0167-8760(94)90027-2. https://www.sciencedirect.com/science/article/pii/0167876094900272.

[14] Bogdan Vlasenko, David Philippou-Hübner, Dmytro Prylipko, Ronald Böck, Ingo Siegert, and Andreas Wendemuth. 2011. Vowels formants analysis allows straightforward detection of high arousal emotions. In *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.

[15] Mohit Shah, Ming Tu, Visar Berisha, Chaitali Chakrabarti, and Andreas Spanias. 2019. Articulation constrained learning with application to speech emotion recognition. *EURASIP journal on audio, speech, and music processing*, 2019, 1, 1–17.

[16] Bjorn Schuller, Bogdan Vlasenko, Dejan Arsic, Gerhard Rigoll, and Andreas Wendemuth. 2008. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. In *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 1333–1336.

[17] Jiahong Yuan, Xingyu Cai, Renjie Zheng, Liang Huang, and Kenneth Church. 2021. The role of phonetic units in speech emotion recognition. *arXiv preprint arXiv:2108.01132*.

[18] Jeng-Lin Li, Tzu-Yun Huang, Chun-Min Chang, and Chi-Chun Lee. 2020. A waveform-feature dual branch acoustic embedding network for emotion recognition. *Frontiers in Computer Science*, 2. ISSN: 2624-9898. DOI: 10.3389/fcomp.2020.00013. https://www.frontiersin.org/article/10.3389/fcomp.2020.00013.

[19] N Cummins et al. 2020. A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. *Proceedngs of Interspeech 2020*, 2182–2186.

[20] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42, 4, 335–359.

[21] Tobias Bocklet, Andreas Maier, Josef G Bauer, Felix Burkhardt, and Elmar Noth. 2008. Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1605–1608.

[22] Rivarol Vergin, Azarshid Farhat, and Douglas O'Shaughnessy. 1996. Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. Volume 2. IEEE, 1081–1084.

[23] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition*, 44, 3, 572–587.

[24] Frans A Boiten. 1998. The effects of emotional behaviour on components of the respiratory cycle. *Biological Psychology*, 49, 1, 29–51. ISSN: 0301-0511. DOI: https://doi.org/10.1016/S0301-0511(98)00025-8. https://www.sciencedirect.com/science/article/pii/S0301051198000258.

[25] Qiang Zhang, Xianxiang Chen, Qingyuan Zhan, Ting Yang, and Shanhong Xia. 2017. Respiration-based emotion recognition with deep learning. *Computers in Industry*, 92-93, 84–90. ISSN: 0166-3615. DOI: https://doi.org/10.1016/j.compind.2017.04.005. https://www.sciencedirect.com/science/article/pii/S0166361516303104.

[26] Ann MacLarnon and Gwen P. Hewitt. 1999. The evolution of human speech: the role of enhanced breathing control. *American journal of physical anthropology*, 109, 341–63. DOI: 10.1002/(SICI)1096-8644(199907)109:3<341::AID-AJPA5>3.0.CO;2-2.

[27] Venkata Srikanth Nallanthighal, Aki Härmä, and Helmer Strik. 2020. Speech breathing estimation using deep learning methods. In *ICASSP 2020 - 2020 IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP), 1140–1144. DOI: 10.1109/ICASSP40776.2020.9053753.

[28] Björn W. Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, Harald Baumeister, Alexis Deighton MacIntyre, and Simone Hantke. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. In *Proc. Interspeech 2020*, 2042–2046. DOI: 10.21437/Interspeech.2020-32.

[29] Maxim Markitantov, Denis Dresvyanskiy, Danila Mamontov, Heysem Kaya, Wolfgang Minker, and Alexey Karpov. 2020. Ensembling End-to-End Deep Models for Computational Paralinguistics Tasks: ComParE 2020 Mask and Breathing Sub-Challenges. In *Proc. Interspeech 2020*, 2072–2076. DOI: 10.21437/Interspeech.2020-2666.

[30] Venkata Srikanth Nallanthighal, Zohreh Mostaani, Aki Härmä, Helmer Strik, and Mathew Magimai-Doss. 2021. Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings. *Neural Networks*, 141, 211–224. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2021.03.029.

[31] Zohreh Mostaani, Venkata Srikanth Nallanthighal, Aki Härmä, Helmer Strik, and Mathew Magimai-Doss. 2021. On the relationship between speech-based breathing signal prediction evaluation measures and breathing parameters estimation. In *Proceedings of ICASSP*, 1345–1349. DOI: 10.1109/ICASSP39728.2021.9414756.

[32] Zohreh Mostaani, RaviShankar Prasad, Bogdan Vlasenko, and Mathew Magimai-Doss. 2022. Modeling of pre-trained neural network embeddings learned from raw waveform for covid-19 infection detection. In *Proceedings of ICASSP*, 8482–8486. DOI: 10.1109/ICASSP43922.2022.9746271.

[33] Zohreh Mostaani and Mathew Magimai.-Doss. 2022. On breathing pattern information in synthetic speech. In *Proceedings of Interspeech*.

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[35] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*. Volume 33, 12449–12460.

[36] Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. 2020. Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters*, 27, 600–604.

[37] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: self-supervised speech representation learning by masked prediction of hidden units. *IEEE Transactions on Audio, Speech, and Language Processing*, 1–1.

[38] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2021. Byol for audio: self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[39] Luyu Wang, Pauline Luc, Yan Wu, Adria Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, et al. 2021. Towards learning universal audio representations. *arXiv preprint arXiv:2111.12124*.

[40] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3875–3879.

[41] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2022. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. *arXiv preprint arXiv:2204.12260*.

[42] Dading Chong, Helin Wang, Peilin Zhou, and Qingcheng Zeng. 2022. Masked spectrogram prediction for self-supervised audio pre-training. *arXiv preprint arXiv:2204.12768*.

[43] Sarthak Yadav and Neil Zeghidour. 2022. Learning neural audio features without supervision. *arXiv preprint arXiv:2203.15519*.

[44] Jort F. Gemmeke, Daniel P. W. 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)*, 776–780.

[45] Mingxing Tan and Quoc Le. 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[46] Sarthak Yadav. 2022. audax. Version 0.x.x. (February 2022). https://github.com/SarthakYadav/audax.

[47] Wei-Ning Hsu et al. 2021. Hubert: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451–3460.

[48] Sanyuan Chen et al. 2021. Wavlm: large-scale self-supervised pre-training for full stack speech processing. *ArXiv*.

[49] Shu-wen Yang et al. 2021. Superb: speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

[50] Dimitri Palaz, Ronan Collobert, and Mathew Magimai-Doss. 2013. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In *INTERSPEECH*.

[51] Tara Sainath, Ron J Weiss, Kevin Wilson, Andrew W Senior, and Oriol Vinyals. 2015. Learning the speech front-end with raw waveform cldnns.

[52] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. 2016. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*.

[53] Hannah Muckenhirn, Mathew Magimai Doss, and Sébastien Marcell. 2018. Towards directly modeling raw speech signal for speaker verification using cnns. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4884–4888.

[54] Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 1021–1028.

[55] Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai-Doss. 2018. On learning to identify genders from raw speech signal using cnns. In *Interspeech*, 287–291.

[56] S Pavankumar Dubagunta, Bogdan Vlasenko, and Mathew Magimai Doss. 2019. Learning voice source related information for depression detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6525–6529.

[57] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. 2021. {leaf}: a learnable frontend for audio classification. In *International Conference on Learning Representations*. https://openreview.net/forum?id=jM76BCb6F9m.

[58] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15, 1, 1929–1958.

[59] 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 Multimodal Sentiment Analysis Challenge*. Workshop held at ACM Multimedia 2022, to appear. Association for Computing Machinery, Lisbon, Portugal.

[60] 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, 76–81.

[61] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore Sound Classification Using Image-Based Deep Spectrum Features. In *Proc. Interspeech 2017*, 3512–3516. DOI: 10.21437/Interspeech.2017-434.

[62] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7.

[63] Tilak Purohit, Imen Ben Mahmoud, Bogdan Vlasenko, and Mathew Magimai Doss. 2022. Comparing supervised and self-supervised embedding for exvo multi-task learning track. *Proceedings of the ICML Expressive Vocalizations Workshop*. Workshop held in conjunction with the 39th International Conference on Machine Learning.

[64] Alice Baird et al. 2022. The icml 2022 expressive vocalizations workshop and competition: recognizing, generating, and personalizing vocal bursts. (2022). DOI: 10.48550/ARXIV.2205.01780.