Fusion of Acoustic and Linguistic Information Using Supervised Autoencoder for Improved Emotion Recognition

Bogdan Vlasenko
Idiap Research Institute
Martigny, Switzerland
bogdan.vlasenko@idiap.ch

RaviShankar Prasad
Idiap Research Institute
Martigny, Switzerland
ravi.prasad@idiap.ch

Mathew Magimai.-Doss
Idiap Research Institute
Martigny, Switzerland
mathew@idiap.ch

ABSTRACT

Automatic recognition of human emotion has a wide range of applications and has always attracted increasing attention. Expressions of human emotions can apparently be identified across different modalities of communication, such as speech, text, mimics, etc. The '*Multimodal Sentiment Analysis in Real-life Media*' (MuSe) 2021 challenge provides an environment to develop new techniques to recognize human emotions or sentiments using multiple modalities (audio, video, and text) over in–the–wild data. The challenge encourages to jointly model the information across audio, video and text modalities, for improving emotion recognition. The present paper describes our attempt towards the MuSe–Sent task in the challenge. The goal of the sub–challenge is to perform turn–level prediction of emotions within the arousal and valence dimensions. In the paper, we investigate different approaches to optimally fuse linguistic and acoustic information for emotion recognition systems. The proposed systems employ features derived from these modalities, and uses different deep learning architectures to explore their cross–dependencies. Wide range of acoustic and linguistic features provided by organizers and recently established acoustic embedding *wav2vec 2.0* are used for modeling the inherent emotions. In this paper we compare discriminative characteristics of hand–crafted and data–driven acoustic features in a context of emotional classification in arousal and valence dimensions. Ensemble based classifiers were compared with advanced supervised autoendcoder (SAE) technique with Bayesian Optimizer hyperparameter tuning approach. Comparison of uni– and bi–modal classification techniques showed that joint modeling of acoustic and linguistic cues could improve classification performance compared to individual modalities. Experimental results show improvement over the proposed baseline system, which focuses on fusion of acoustic and text based information, on the test set evaluation.

CCS CONCEPTS

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

MuSe '21, October 24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8678-4/21/10…$15.00
https://doi.org/10.1145/3475957.3484448

KEYWORDS

Emotion recognition, wav2vec2, Supervised auto encoders, Bag-of-audio-words, late fusion

ACM Reference Format:
Bogdan Vlasenko, RaviShankar Prasad, and Mathew Magimai.-Doss. 2021. Fusion of Acoustic and Linguistic Information Using Supervised Autoencoder for Improved Emotion Recognition. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge (MuSe '21), October 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3475957.3484448

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 their recognition in 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 [1–7]. Speech and text based modeling methods are more popular among all modalities for the task of emotion recognition, as they imbibe prosodic and semantic information, while giving significant improvements [8]. Studies have been performed towards appropriate annotation of words with emotion specific tags in arousal and valence dimensions [9]. Several studies have also been made on sentiment analysis over videos, most of which utilize an automatic transcription pipeline, followed by a concept extraction module, which leads to extraction of emotions [10, 11]. Other studies, focused on fusing biological signals with speech features to model the emotional state, rely on the Long Short–term memory (LSTM)–Recurrent neural network (RNN) model with attention mechanism to model contextual dependencies [12]. Physiological signals have also been employed in an end–to–end framework to derive arousal and valence states using convolutional and recurrent layers [13]. To explore the efficacy of proposed methods, cross–corpora studies have been performed to examine the dependence of factors such as, emotion type, normalization methods, languages, and speakers, towards emotion recognition [14]. Considering significant results with combining the linguistic and acoustic information towards the task of depression detection [15, 16], the present study explores different ways

of fusion of speech and textual information for categorical sensing emotions 'in the wild' scenarios.

Multimodal Sentiment Analysis (MuSe) 2021 challenge [17] motivates towards integrating knowledge across audio–visual signals, text, and physiology disciplines to attempt the challenges in emotion recognition using multiple modalities. The challenge poses the tasks of classification and regression towards emotion, physiological emotion and stress and sentiment recognition. The current paper targets the task Multi–modal Sentiment in–the–wild classification (MuSe–Sent) sub–challenge to recognise emotions in the arousal and valence dimensions. The challenge introduces a gold standard fusion method, Rater Aligned Annotation Weighting (RAAW), for continuous annotations, to improve the inter–rater agreement and minimize the variance in annotator reaction times. The challenge encourages unification of disciplines via fusion and exploit the co–dependencies across different modalities. The challenge gives late fusion results, obtained by aggregating the predictions from different models trained individually on different feature set. The baseline systems employ a LSTM–RNN based architecture to exploit the sequential information in the features. The LSTM–RNN network is followed by a feed–forward layer which appropriately gives sequence of logits, or single–valued prediction for regression and classification tasks, respectively. Open source tools are used to derive a wide–range of feature set for building a variety of baseline systems. The features obtained across different modalities are aligned with their labels using the Montreal Forced Aligner (MFA) tool.

In this work, we build multiple systems for predicting emotional state in the arousal and valence dimensions from linguistic and acoustic cues. In addition to baseline feature provided by organizers we utilize advanced *wav2vec2* embedding, popularly used in state–of–the–art ASR systems. Also, in order to map sequential data into fixed–length acoustic feature vectors, we use Bag-of-Audio-Words for acoustic feature modeling. We compare the modeling of emotions using these features based on ensemble classifier and supervised autoencoder (SAE).

Our major contributions to the challenge in this paper are as follows:

- compare *hand–crafted* features and *data driven* DNN–based acoustic embeddings in a context of acoustic information based emotion recognition
- evaluating multiple feature representation techniques for mapping *frame–level* to *turn–level* feature space
- compare *ensemble classification methods* with supervised auto encoder (SAE) technique for emotion–recognition with turn–level features
- selection of the most *robust uni–modeling* techniques for the proposed feature space
- evaluating *early–* and *late–fusion* techniques for boosting uni–modal classification techniques, combining evidences from linguistic and acoustic based information.

The rest of the paper is organized as follows. Sec. 2 describes the feature representations derived from the linguistic and acoustic modalities, utilized in the subsequent studies. The section also describes the methods used to create different modules in the proposed systems. Sec. 3 explains the dataset, evaluation metrics, methods utilized to create proposed systems, and the experimental setup of the baseline and proposed systems. Sec. 4 gives the results obtained using different methods and systems. Sec. 5 gives a conclusion to the paper.

2 METHODS

The section describes the acoustic and linguistic feature set, derived from audio signals and the corresponding text transcriptions, respectively. The section further discusses machine learning architectures and principles, given in the baseline paper as well as proposed in the current study, towards the task of emotion recognition.

2.1 Features representation

2.1.1 Acoustic features. For modeling the acoustic information for emotion classification in arousal and valence dimensions, we decided to utilize hand crafted features provided with openSmile toolkit [18] and state–of–the–art acoustic embeddings.

LLDs and LLDs x functionals: a set of hand crafted features extracted on frame and turn level. Low–level descriptors (LLDs) obtained from the signal, based on the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [19], are extracted using the open–source openSmile toolkit [18]. The parameter set eGeMAPS comprises of 88 LLDs characterizing the frequency, energy, and spectral and temporal behavior, of the signal. It is a minimal set of hand-crafted features, reflecting on physiological changes in voice production, and has proven effective for automatic voice analysis tasks [20]. The eGeMAPS parameters represent frame–level features, and additional mapping is required to extract segment level features. Along with this, we use the ComParE [21] hand crafted turn–level feature set. This set comprises of 6373 static features resulting from the computation of functionals (statistics) over low–level descriptor (LLD) contours.

Wav2vec2 features: we investigated framework for deriving feature set based on self-supervised learning of representations from raw audio data. The embedding framework has shown significant potential while pre–training on unlabeled data for speech recognition systems [22]. Recent studies have shown that *wav2vec2* embeddings can effectively be used for robust emotion recognition from speech [23].

2.1.2 Linguistic features. The semantic information in the available transcriptions for the audio information, is captured using Bert features.

Bert: set of natural language processing (NLP) features, to map the sequential information in the transcriptions provided with the challenge. Transformer language model, namely Bidirectional Encoder Representations from Transformers (BERT) [24], which have already been successfully used for a variety of NLP tasks, are used to learn semantic representations from the text. BERT pre–trains its deep representations on context of unlabelled text, and further fine–tunes them on a broad selection of down–streaming NLP tasks. The context–based representations are preserved, while yielding one feature vector per word. The features are derived from the last four BERT layers resulting in a 768 dimensional feature vector analogous to the study in [25]. This is in contrast to static word embeddings which give one vector per word independent of the

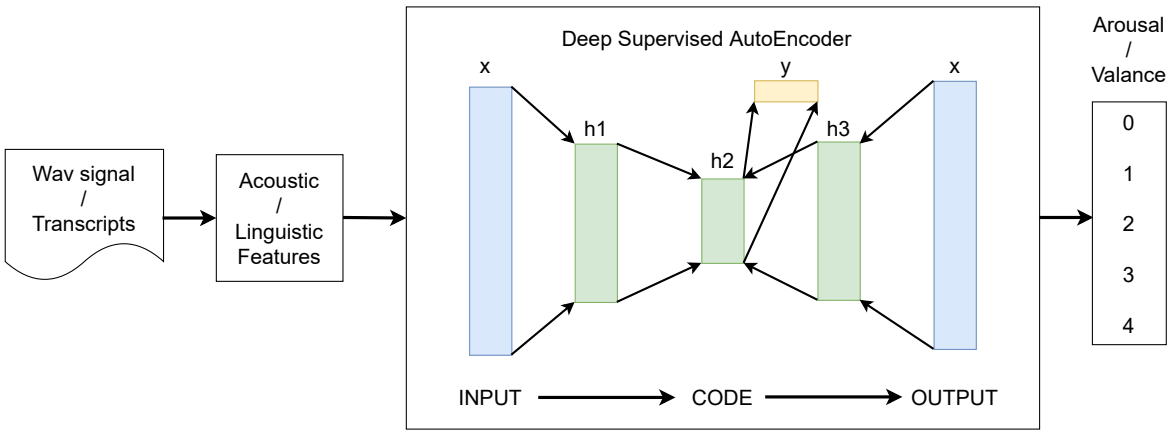

Figure 1: The supervised autoencoder (SAE) network architecture. Acoustic or linguistic features are used as input to the SAE. The target "y" is used to determine the supervised loss. The classification layer outputs the arousal/valence levels.

context. For **MuSe–Sent**, the base variant of BERT, pre-trained on English texts, is used.

2.1.3 Turn-level features generation. Considering that most features used in our study incorporate a frame–level information we decided to employ several mapping techniques for generating turn–level features

Bag-of-Audio-Words (BoAW): In order to capture the sequential information into fixed length feature vectors we used the BoAW approach. These features have successfully been applied for various speech applications such as, acoustic event detection and speech–based emotion recognition [26, 27]. Audio chunks are represented as histograms of acoustic LLDs, after quantization based on a codebook. In our experimental study we used 100 and 50 samples per codebook and *kmean++* clustering for codebook generation. The openXBOW toolkit is used for deriving BoAW representation [28].

Functional mapping For mapping word level BERT embeddings onto turn–level segments we decided to use a short list of functionals. We applied 3 functional: *average*, *mean* and *standard deviation* to map sequential information into fixed–length representation. Hence a $768 \times 3 = 2304$ dimensional vector is used as feature vector for NLP based techniques, henceforth referred as to BERT funct.

2.2 Machine learning techniques

To select the most robust machine learning technique for emotion classification we used set of state–of–the–art ensemble classification methods: *RandomForest, AdaBoost* and *GradientBoosting*. The main goal of ensemble classification is to optimally weigh the predictions of multiple base estimators, realised with different architectures and learning algorithms, in order to improve generalizability/robustness over a single estimator. As alternative machine learning technique we utilize the supervised autoencoder network, which has shown considerable generalization abilities for natural language processing tasks [29]. The overall architecture of the proposed SAE is shown in Fig. 1. The following subsections briefly describes the components of training the SAE.

2.2.1 Supervised Autoencoder (SAE). SAE is an autoencoder (AE) network where a supervised loss is imposed across the representation layer [30]. For networks with single hidden layer, the supervised loss operates upon the output layer, while for deeper AEs, the supervised loss component is added to the bottleneck or the innermost layer. A combination of the supervised loss and the reconstruction loss effectively captures the underlying patterns in the input data, along with improving upon modeling accuracy of the input representation. The representations are learned in a lower dimensional space k from input $\mathbf{x}$ with dimension d using the transformation $\mathbf{F} \epsilon \mathbb{R}^{d \times k}$. The reconstruction learns an inverse transformation $W_r \epsilon \mathbb{R}^{k \times d}$ along with the transform $W_p \epsilon \mathbb{R}^{k \times m}$ to predict the target y. The SAE network combines the reconstruction loss $L_r(W_r, Fx_i, x_i)$ (where x_i is training sample) and the supervised loss $L_p(W_p, Fx_i, y_i)$ (where y_i is target) to improve generalization while minimizing the reconstruction error. The objective function is given as

$$\frac{1}{t}\sum_{i=1}^{t}\left[L_p(W_p, Fx_i, y_i) + L_r(W_r, Fx_i, x_i)\right] = \frac{1}{t}\sum_{i=1}^{t}\left[\|W_p Fx_i - y_i\|_2^2 + \|W_r Fx_i - x_i\|_2^2\right]. \quad (1)$$

The supervised loss aims to focus the representation learning towards task oriented representations. SAE has successfully been employed for several engineering tasks, such as image encoding and classification, and language identification [29, 31].

2.3 Fusion techniques

In order to find appropriate combination of acoustic and linguistic cues processing for robust bi–modal emotion recognition early (*EF*) and late fusion (*LF*) techniques were used. In the case of *LF* techniques, weighted majority voting (*WMA*) with two different implementations:

- **simple weights concept** LF_{SW}: posteriors obtained with two different classification techniques were averaged, and the class with highest posterior is selected as predicted output.
- **combined weights concept** LF_{CW}: posteriors obtained with two different classification techniques are weighted and averaged. Weights are determined by class–level classification performance in terms of $f1$ score. The highest weighted posterior is selected as predicted output.

Hence, prior knowledge bout the prediction performance of the classifier is needed the second concept have been used for the evaluations on the test set.

3 EXPERIMENTAL SETUP

The section introduces the dataset and evaluation metric used for the study. A brief description of the baseline systems provided with the challenge is also presented in the section. A further description of the proposed methods and techniques is also presented.

3.1 Dataset

The MuSe–Sent sub–challenge uses the Multimodal Sentiment Analysis in Car Reviews (**MuSe–CAR**) dataset which contains almost 40 hours of multimodal data, annotated with the corresponding emotions. The dataset was created in–the–wild with the goal of Multimodal Sentiment Analysis. The dataset comprises of a selection of videos from *YouTube* along with metadata information about the speaker's age, nativity, dialect, camera shot range and angles, scene settings, and additional noise and settings. The dataset is automatically transcribed at word level using the Google Cloud speech API, and Amazon Transcribe, for verbal and non–verbal (laughter, music, etc.) elements. The transcriptions include the durational timestamps for words, and appropriate punctuation. Further annotation of the data for the emotion dimensions, speaker characteristics, and topics and entities, is performed manually. Software tools, such as ELAN and DARMA are used to annotate multimodal data for categorical information, and continuous emotions, respectively. The challenge requires to predict among 5 sentiment classes for the valence and arousal emotion dimensions, at a segment level. The continuous dimensional are mapped to categorical representations for emotion for either dimensions based on the unsupervised clustering of time–series features extracted from the input data. Separate clustering techniques were applied for arousal and valence dimensions. The classes are identified using numeric labels in a range 0..4 used in MuSe–Sent, and do not correspond to the degree of emotional expression in arousal or valence dimensions, but to the cluster centroids. The ground truth is created by mapping a continuous space of annotated emotions to a categorical representation, via unsupervised clustering over time–series features obtained over segments of data. Tab. 1 gives the partitions of data utilized for training.

Table 1: Distribution of instances across training and development sets, in valence and arousal dimensions for the MuSe–Sent sub–challenge, for different classes (cl)

Arousal				Valence			
cl	Train	Devel	Test	cl	Train	Devel	Test
0	612	249	–	**0**	528	71	–
1	534	135	–	**1**	552	159	–
2	312	96	–	**2**	1178	458	–
3	1255	388	–	**3**	1112	405	–
4	1494	467	–	**4**	837	242	–
Σ	**4207**	**1335**	**1260**	Σ	**4207**	**1335**	**1260**

3.2 Evaluation metrics

The classification task **MuSe–Sent** is evaluated in terms of *f1* score (macro), an evaluation measure commonly used for class–imbalanced data. The ultimate goal of the challenge is to reach the highest possible average $f1$ score over all 5 classes.

3.3 Baseline systems

The challenge provides baseline systems which can be used to evaluate the classification performance of features and networks for emotion dimensions. Time–series acoustic features are derived at segment level from the input data. These features are further normalized, and transformed to a denser space using PCA, to obtain 5 clusters. Single channel audio is extracted from the video recordings, and acoustic features are extracted using openSMILE and DeepSpectrum tools. For acoustic data, handcrafted features (eGeMAPS acoustic parameter set, comprising of 88 dimensional features), deep features (4096 dimensional DeepSpectrum features, based on convolutional neural networks (CNNs)), and VGGish features (128 dimensional embeddings, obtained from log spectrograms), are derived. Textual features, derived over the annotations in the dataset using a Transformer language model, are also used in contrast to acoustic features. The BERT method, pretrained on context of unlabelled text, is used to derive a 768 dimensional feature vector. The acoustic and linguistic features are fused in an early/late manner to achieve a desirable classification across valence and arousal emotion dimensions.

The baseline systems are designed using a Long Short–Term Memory (LSTM)–RNN based architecture, with a hidden state dimensionality within the range $h = \{32, 64, 128\}$. Different systems with LSTM–RNN layers (n) varying in $n = \{1, 2, 4\}$ are trained. The learning rate is varied across $lr = \{0.001, 0.005, 0.01\}$. The sequence obtained from the final LSTM–RNN layer is connected to a feed–forward layer to obtain a target classification label. The features are aligned with the labels and are provided as a part of the *MuSe–Car* sub–challenge. Repetition of features over a duration length and zero–padding to frames with unavailable features is appropriately done. The organizers mentioned that the best possible score ($f1 = 38.27$) for emotional arousal prediction could be obtained just with deep representation BERT features, which captures context–based representation per word (for more details see Tab 2).

Table 2: $f1$ scores obtained on the development set for uni-modal techniques and baseline systems. Abbreviations: rb-reported in baseline paper, o-obtained, Aro - arousal, Val - valence, Comb - combined

MuSe–Sent				
Features	**Classifier**	**Aro**	**Val**	**Comb**
Baseline systems				
DeepSpectrum(rb)	LSTM	33.5	30.2	31.9
DeepSpectrum(o)	LSTM	23.2	19.2	21.2
eGeMAPS(rb)	LSTM	36.0	32.9	34.5
eGeMAPS(o)	LSTM	20.2	22.0	21.1
Bert(rb)	LSTM	38.3	32.7	35.5
Bert(o)	LSTM	23.3	16.3	19.7
Proposed systems				
BERT funct	ensemble	31.2	31.9	31.5
BoAW(Bert)	ensemble	33.1	30.3	31.7
BoAW(eGeMAPS)	ensemble	33.9	30.1	32
BoAW(Wav2vec2)	ensemble	32.7	31.1	31.9
BoAW(eGeMAPS)	SAE	34.1	32.6	33.4
BoAW(Wav2vec2)	SAE	32.7	31.1	31.9
BoAW(Bert)+ BERT funct	SAE	35.1	30.8	32.9
ComPaRE	SAE	35.3	31.2	33.2

3.4 Proposed systems

To improve upon the DNN based baseline systems in the baseline paper, we experiment with a variety of acoustic features and classification techniques. While attempting towards MuSe–Sent task, we focus on using turn–level features for emotion recognition. Hence, during first stage of our experimental study we extended the set of acoustic features obtained by data–driven LLDs, with *wav2vec2*– and turn–level functional feature set(ComParE). Afterwards, we implemented *BoAW* and functional mapping to the linguistic and acoustic features to derive turn–level feature representation. Two different configurations of BoAW codebook based representations were evaluated: 50 codebook vectors extracted with random sampling approach, and 100 codebook vectors generated with *k–mean++*. For mapping word–level BERT feature into turn–level representation in addition to BoAW concept we used functional mapping concept.

The next challenge for us to select an appropriate machine learning technique for processing turn-level features. We chose to compare the results obtained with a set of ensemble based classifiers on one hand, and supervised autoencoder on the other. To select the most robust ensemble classification techniques, the grid search concept integrated in *sklearn* package was used. Tuning of hyper-parameters for ensemble based classifiers was conducted within 10 fold cross–validation experiments on the training set. The best configuration of the ensemble based classifiers was selected for experiments conducted over partitions for training and development sets, given by organizers.

The basic SAE architecture described in Fig. 1, is implemented in PyTorch. Considering sufficient number of training samples, up to 1000 epochs were used to tune the network parameters. To optimize the SAE architecture we optimize from a range of hyper parameters as follows. The number of hidden layers: 1–5, learning rate: 10^{-5}–10^{-2}, weight decay: 10^{-6}–10^{-3} and activation functions: 'relu', 'sigma'. Optimization of the SAE architecture was conducted during training for each input feature set. We assume, that SAE based classification techniques could provide a more stable classification performance for in-balanced data sets like MuSe 2021 Sent challenge.

3.4.1 Bayesian Optimizer (BO) . In the case of SAE, there are many hyperparameters related to model design and optimization. AE training and performance often benefit from hyperparameter tuning to avoid over– and under–fitting. BO is a state-of-the-art hyperparameter optimization algorithm which has achieved competitive performance on several optimizations benchmarks [32, 33]. BO is a technique based on Bayes theorem, and works by building a probabilistic model of the objective function, called the surrogate function. This function is then searched efficiently with an acquisition function before candidate samples are chosen for evaluation on the actual objective function.

Considering wide diversion of feature representation techniques we also applied early fusion for acoustic and linguistic features during selection of the most robust uni-modal classification technique. Posteriors obtained with the most robust uni-modal classifiers were used for late fusion of concepts introduced earlier.

4 EXPERIMENTAL STUDY

The section describes the results obtained using baseline systems provided with the challenge. The section further discusses the confusion matrices obtained using the proposed features and networks. Considering the fact that the cardinality of classes is not related to the emotional degree, the vicinity of classes in the confusion matrices do not reflect close emotional level. The section finally describes the $f1$ scores on test dataset, obtained using fusion of best techniques.

4.1 Baseline systems

We implemented the baseline systems provided with the challenge using the suggested protocol. It is to be noted that the baseline results reported by the challenge paper couldn't be recreated with the given systems. We attempted to follow the experimental protocol as guided by the challenge paper, but found different scores. Tab. 2 gives the $f1$ scores obtained using the baseline systems (o), along with the scores reported in the baseline paper (rb), using the given features and systems.

4.2 Uni–modal systems

We further trained and evaluated our methods for the MuSe–Sent sub–challenge training and development datasets. For our experiments, we focused on features from acoustic and linguistic modalities. Following the train–development partitioning that has been obtained over the original data set, we processed 4207 samples for training, and 1335 samples for validation.

During first experimental phase we compared *k–mean++* (100 vectors codebook) and random sampling (50 vectors codebook) for turn–level BoAW representation with ensemble based classification

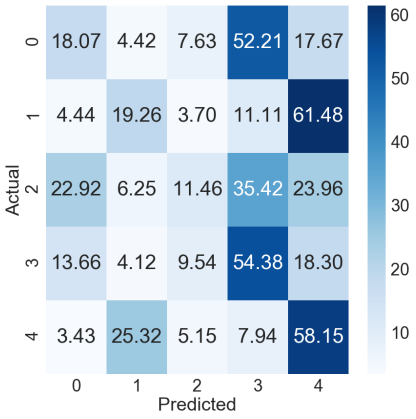

Figure 2: Confusion matrix for arousal prediction on the development set for BERT FUNCT using AdaBoost classifier.

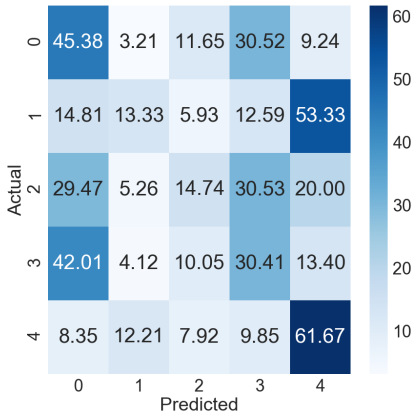

Figure 3: Confusion matrix for arousal prediction on the development set for BoAW(BERT) using AdaBoost classifier.

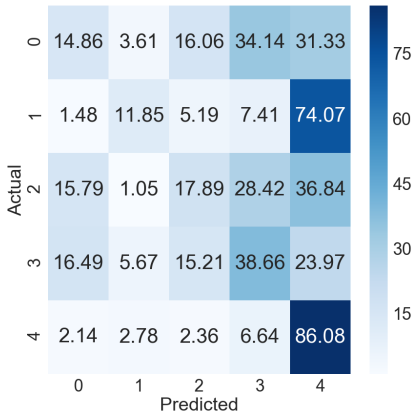

Figure 4: Confusion matrix for arousal prediction on the development set for BoAW(EGEMAPS) using AdaBoost classifier.

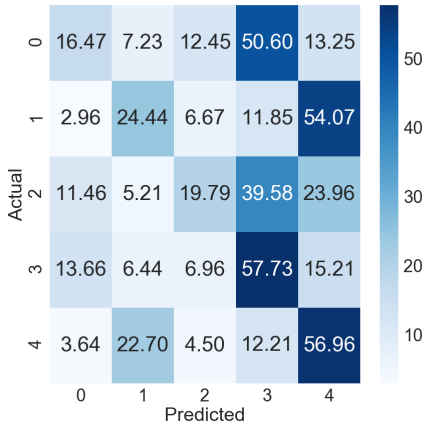

Figure 5: Confusion matrix for arousal prediction on the development set for *EF* of acoustic and linguistic features using SAE classifier

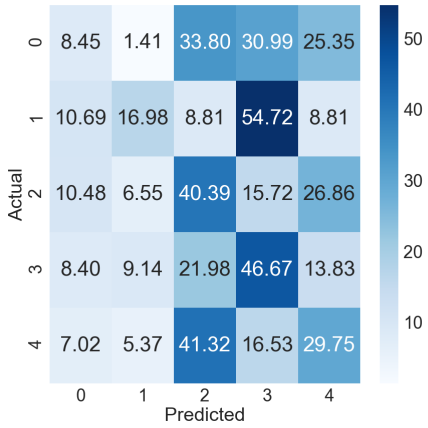

Figure 6: Confusion matrix for valence prediction on the development set for *EF* of acoustic and linguistic features using SAE classifier

techniques. We find out that *k–mean++* based BoAW provides better turn level representation. Hence, in the following experimental study we decided to use 100 dimensional BoAW feature vector for BERT, EGEMAPS and WAV2VEC2 feature representations. During the second experimental phase, we find out that hand–crafted features outperforms data–driven DNN based *wav2vec2* features, over both ensemble and SAE based techniques. Hence we decided to proceed with hand–crafted features for the proposed acoustic based system evaluated on test set.

We performed emotion classification for arousal and valence dimension using linguistic features derived using BERT, and BERT FUNCT, and acoustic features derived using EGEMAPS, over the MUSE–SENT development set. Fig. 2 gives the confusion matrix obtained using BERT FUNCT features for prediction of emotion classes in arousal dimension. A comparison of confusion matrix obtained with BoAW representation derived over BERT (Fig. 3) shows that the corresponding fixed length representation improves upon the weightage across the diagonal of the confusion matrix.

Table 3: $f1$ scores obtained on the development set for early- and late-fusion with combined audio and text analysis.

		MuSe–Sent						
Conf#	**Fusion**	**Audio analysis**		**Text analysis**		**Arousal**	**Valence**	**Combined**
		Features	**Classifier**	**Features**	**Classifier**			
1	LF_{SW}	BoAW(eGeMAPS)	ensemble	BoAW(BERT)+BERT funct	SAE	36.3	31.5	33.9
2	*EF*	ComParE	SAE	BERT funct	SAE	35.2	31.4	33.3
3	*EF*	BoAW(eGeMAPS)	SAE	BoAW(BERT)+BERT funct	SAE	36.8	30.9	33.9
4	LF_{SW}	BoAW(eGeMAPS)	ensamble	BoAW(BERT)+BERT funct	SAE	36.5	31.8	34.2
5	LF_{SW}	ComParE	SAE	BERT funct	SAE	35.4	31.7	33.6
6	LF_{SW}	BoAW(eGeMAPS)	SAE	BoAW(BERT)+BERT funct	SAE	36.9	31.3	34.1

The BoAW based representation, however, does lead to increased confusion between few classes, but overall results in higher *f1*. In order to combine discriminative characteristics of the BERT and BERT funct we decided to combine those two types of linguistic features representation with early fusion (EF) approach. Hence, results presented in Tab. 2 proves that combination of BoAW and Functional based representation could improve overall $f1$ score for text based emotion recognition system. Confusion matrix for the BoAW(eGeMAPS) (Fig. 4) shows that pure acoustic modeling with representations obtained using eGeMAPS features could not provide a good average *f1* score distributed over all 5 classes.

Also, considering results presented in the Tab 2 one could see that SAE approach provides better distribution of class–level $f1$ scores in comparison with ensemble based techniques. Based on results presented in Tab 2 we selected the set of the most robust uni-modal classification techniques namely:

- audio information representation based on ComParE set, BoAW(eGeMAPS) with SAE and Adaboost (ensemble) classifiers,
- linguistic information representation using BoAW(BERT) with SAE classifier, and early fusion of BoAW(BERT) + BERT funct with SAE.

Expressions of human emotions can apparently be identified across different modalities of communication, such as speech, text, mimics. In this talk, I'll describe methods used for fusion linguistic and acoustic information during participation in 'Multimodal Sentiment Analysis in Real-life Media' (MuSe) 2021 challenge. In our experimental study, just linguistic and acoustic information was used. Selection of optimal suprasegmental features representation, machine learning techniques and fusion techniques will be presented. Overview of knowledge-based acoustic feature representation and corresponding toolkits will be provided. Finally, the proposed multimodal system with fused acoustic and linguistic information channels achieved 3rd place for Arousal and Valence 18th (out of 60 participants and 181 submissions).

4.3 Bi–modal systems

For improving upon the performances with baseline systems (Section 3.3), and uni–modal systems (Section 4.2), we experimented with joint modeling of acoustic and linguistic modalities. A balanced fusion of performances over these modalities is expected to yield even weightage across the diagonal of confusion matrices. Early fusion (*EF*) for openSmile and BERT features have been used to explore the ability of bi–modal modelling towards overall classification performance. Confusion matrices for *EF* based system with openSmile and BERT features are presented in Figs. 5 and 6. As can be noted from the figures, early fusion of acoustic and linguistic features could improve diagonal representation for overall $f1$ score, compared to modeling only the acoustic information (Fig 4).

As a final step of out experimental study, we applied LF_{SW} in pairwise manner to top performing uni–modal modeling techniques. Posteriors probabilities generated with SAE based classifier and ensemble based technique were used to determine LF_{SW} and LF_{CW} components, and hence the predicted output.

The *f1* scores obtained using the *LF* techniques for bi–modal modeling on the development set are given in Tab. 3. The scores illustrate significant improvement *f1* performance with the joint–modeling of different modalities. Among the classifiers used, SAE gives better performance in discriminating among 5 classes, as compared to ensemble classifiers. Finally, considering the results obtained with bi–modal fusion of good performing linguistic and acoustic based emotion classifiers, we chose the techniques giving highest $f1$ scores for test set evaluation. During final trial for MuSe 2021 Sent challenge we use late fusion technique with weighted average posteriors. The systems chosen for evaluation of test set are given in Tab. 3.

As noted from the Table 4 the best average $f1$ score for arousal prediction was obtained with LF_{SW} for bi-modal modeling with supervised auto encoder and BoAW(eGeMAPS) and combined NLP feature set representation with BoAW(BERT) and BERT Funct. For valence dimentionality prediction the best performance was obtained with lafe fusion technique with complex weights approach.

5 CONCLUSIONS

In this paper, we studied multiple systems for predicting emotional state in the arousal and valence dimensions using linguistic and acoustic information. We extended investigations with acoustic feature representations with *wav2vec2* and hand–crafted features (ComParE). Our experimental results showed that *wav2vec2* acoustic embeddings do not provide any noticeable gain over the hand–crafted features for emotion specific information. In further experiments, we experimented with uni–modal systems by using an ensemble of classifiers to obtain improvement in performance. Bi–modal systems are further experimented with the top performing uni–modal factors. We illustrated that the functional mapping

Table 4: $f1$ scores obtained on the test set with optimized configurations. Abbreviations: A - audio, T - text, LF late fusion, SW simple weights, CW combined weights

MuSe–Sent				
Fusion	**Config**	**Valence**	**Arousal**	**Combined**
Baseline				
best LF A+T	-	30.29	32.87	31.6
Bert	-	31.90	30.63	31.27
DeepSpectrum	-	27.26	33.16	30.21
eGeMAPS	-	25.80	31.97	28.89
Proposed				
LF_{SW}	6	27.92	**33.68**	30.8
LF_{SW}	3	25.04	30.85	27.94
LF_{SW}	3	26.75	32.78	29.77
LF_{SW}	4	24.16	28.92	26.54
LF_{CW}	6	**28.01**	32.53	30.27

concept in combination with bag-of-audio-words representation improves the NLP based classification performance. With our experiments over classifiers, we observed that supervised autoencoder provides better generalization over acoustic and linguistic feature space, given an unbalance of class instances in MuSe 2021 data set.

Bi–modal approaches are devised with fusing the best performing systems in the uni–modal approach, which lead to improvement in performance. Presented experimental results show that combining acoustic and linguistic information processed with BoAW approach captures significant amount information related to arousal and valence emotional dimensions. With the proposed late fusion technique we were able to outperform the performance of baseline (fusion of text and audio information) systems for arousal and valence dimensions. The optimal system ranked 3^{rd} among all the participants for the task of arousal state classification. However, performance for emotion state classification in the valence dimension couldn't reach a benchmark. The submission was placed rank 18, with the overall rank 16 among all entries. We are planning to continue our work on more advanced audio and NLP based techniques in order to improve classification performance for valence emotional modality. Hence further studies are being planned to improve the techniques to give a good combined performance.

6 ACKNOWLEDGEMENTS

This work was partially funded by the Innosuisse project CMM: Conversational Member Match (grant no. 38843.1 $IP - ICT$), the Swiss National Science Foundation project TIPS: Towards Integrated processing of Physiological and Speech signals (grant no. 200021_188754) and 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. $40B2 - 0\_194794$). The authors would like to thank Dr. Pavankumar Dubagunta for helping with the wav2vec2 features extraction codes. The authors would also like to thank Tilak Purohit for constructive inputs and suggestions.

REFERENCES

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

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

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

[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. 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*, pages 35–44, 2020.

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

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

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

[8] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. IEEE, 2018.

[9] Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45(1):169–177, 2013.

[10] Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[11] Lukas Stappen, Alice Baird, Erik Cambria, and Bjőrn W Schuller. Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems*, 36(2):88–95, 2021.

[12] Alice Baird, Shahin Amiriparian, Manuel Milling, and Bjőrn W Schuller. Emotion recognition in public speaking scenarios utilising an lstm-rnn approach with attention. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 397–402. IEEE, 2021.

[13] Gil Keren, Tobias Kirschstein, Erik Marchi, Fabien Ringeval, and Bjőrn Schuller. End-to-end learning for dimensional emotion recognition from physiological signals. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 985–990. IEEE, 2017.

[14] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, André Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, July-December 2010.

[15] D S Pavan Kumar, Bogdan Vlasenko, and Mathew Magimai.-Doss. Learning voice source related information for depression detection. In *Proceedings ICASSP 2019,*, pages 6525–6529, 2019.

[16] Esaú Villatoro-Tello, P Dubagunta, Julian Fritsch, G Ramirez-de-la Rosa, Petr Motlicek, and M Magimai-Doss. Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition. INTERSPEECH, 2021.

[17] Lukas Stappen, Alice Baird, Lea Schumann, and Bjőrn Schuller. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing*, -(-):–, 2021.

[18] Florian Eyben, Martin Wöllmer, and Bjőrn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[19] Florian Eyben, Klaus R Scherer, Bjőrn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

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

[21] Bjorn W. Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, Sandra Ottl, Maurice Gerczuk, Panaguiotis Tzirakis, Chloë Brown, Jagmohan

Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, M.Rothkrantz Leon J. Joeri Zwerts, Jelle Treep, and Casper Kaandorp. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In *Proceedings INTERSPEECH 2021*, Brno, Czechia, September 2021. ISCA.

[22] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[23] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Stroudsburg, PA, USA, 2019.

[25] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. Multi-modal 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*, MuSe'20, page 27–34, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381574. doi: 10.1145/3423327.3423672.

[26] Hyungjun Lim, Myung Jong Kim, and Hoirin Kim. Robust sound event classification using lbp-hog based bag-of-audio-words https://www.overleaf.com/project/60d5fce79adf88e6d0a5aa1ffeature representation. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[27] Maximilian Schmitt, Fabien Ringeval, and Bjőrn W Schuller. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *Interspeech*, pages 495–499, 2016.

[28] Maximilian Schmitt and Bjőrn Schuller. Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit. 2017.

[29] Shantipriya Parida, Esaú Villatoro-Tello, Sajit Kumar, Petr Motlicek, and Qingran Zhan. Idiap submission to swiss-german language detection shared task. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, number CONF. CEUR Workshop Proceedings, 2020.

[30] Lei Le, Andrew Patterson, and Martha White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Advances in Neural Information Processing Systems*, pages 107–117, 2018.

[31] Qiuyu Zhu and Ruixin Zhang. A classification supervised auto-encoder based on predefined evenly-distributed class centroids. *arXiv preprint arXiv:1902.00220*, 2019.

[32] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[33] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.