Unimodal Multi-Task Fusion for Emotional Mimicry Intensity Prediction

Tobias Hallmen
Chair for Human-Centered Artificial Intelligence
University of Augsburg
tobias.hallmen@uni-a.de

Fabian Deuser
Institute for Distributed Intelligent Systems
University of the Bundeswehr Munich
fabian.deuser@unibw.de

Norbert Oswald
Institute for Distributed Intelligent Systems
University of the Bundeswehr Munich
norbert.oswald@unibw.de

Elisabeth André
Chair for Human-Centered Artificial Intelligence
University of Augsburg
elisabeth.andre@uni-a.de

Abstract

In this research, we introduce a novel methodology for assessing Emotional Mimicry Intensity (EMI) as part of the 6th Workshop and Competition on Affective Behavior Analysis in-the-wild. Our methodology utilises the Wav2Vec 2.0 architecture, which has been pre-trained on an extensive podcast dataset, to capture a wide array of audio features that include both linguistic and paralinguistic components. We refine our feature extraction process by employing a fusion technique that combines individual features with a global mean vector, thereby embedding a broader contextual understanding into our analysis. A key aspect of our approach is the multi-task fusion strategy that not only leverages these features but also incorporates a pre-trained Valence-Arousal-Dominance (VAD) model. This integration is designed to refine emotion intensity prediction by concurrently processing multiple emotional dimensions, thereby embedding a richer contextual understanding into our framework. For the temporal analysis of audio data, our feature fusion process utilises a Long Short-Term Memory (LSTM) network. This approach, which relies solely on the provided audio data, shows marked advancements over the existing baseline, offering a more comprehensive understanding of emotional mimicry in naturalistic settings, achieving the second place in the EMI challenge.

1. Introduction

Emotional mimicry is a phenomenon where individuals imitate the emotional expressions of others, such as their facial expressions, vocal tones, or body language [16–18]. This mirroring can facilitate social bonding by helping individuals understand and empathise with the emotional states of those around them [17]. For example, when someone smiles, others around them are likely to do the same, creating a shared emotional experience. One particular interesting aspect that this mimicry phenomenon mostly appears when people already have bonds to each other [29]. In therapeutic settings, emotional mimicry can be particularly beneficial, as it helps therapists connect with their clients, making them feel understood and supported.

Leveraging deep learning to discern and anticipate emotional states in therapeutic settings enhances therapists' insight into and reaction to clients' emotions. Utilising indicators like facial expressions, and variations in voice pitch and tone aids in this process. Previous research [7] introduced a multi-modal dataset categorising emotional mimicry into "Approval", "Disappointment", and "Uncertainty". Contemporary studies [11, 14, 43] employ diverse multimodal features from this dataset, including audio-visual and, with the aid of ASR models [36], textual data, further enhanced by textual embeddings [8, 10, 46]. Integrating large language model features with audio-visual data for predicting emotional mimicry brings considerable computational challenges. This is due to the intricate process of combining and analysing different types of data. Therefore, creating a cohesive end-to-end pipeline becomes an important goal. Such a pipeline would provide a harmonious balance between efficiency and performance when handling diverse data streams.

For this study, we participated in the 6th Workshop and Competition on Affective Behavior Analysis in-the-wild, specifically the Emotional Mimicry Challenge. This segment of the challenges focuses on predicting the intensity of mimicked emotions, a crucial aspect that could enhance the precision of therapeutic applications. The challenge utilises video data showcasing individuals mimicking specific seed emotions, with a notable twist: the annotators were unaware of the intended emotions of the seeds, providing intensity

ratings based on their mimicry and seeded emotions. This approach allowed for a more authentic assessment of emotional mimicry intensity, offering valuable insights into how emotions are conveyed and perceived in a naturalistic setting. Compared to the previous dataset [7] it contains a more fine-grained categorisation of the emotions, which is more challenging to separate.

We present an innovative method focused solely on audio to predict emotional mimicry in response to perceived videos. This approach stems from our preliminary findings, which indicated that relying on visual data did not yield satisfactory outcomes and imposed significant hardware limitations during the training process. By exclusively utilising audio data, we conducted evaluations of the given challenge dataset and achieved competitive results. Our findings underscore the critical role of task-specific pre-trained weights and highlight the significance of our architectural decisions in the success of this method.

2. Related work

Within the domain of affective computing, notable advancements have been achieved in a variety of tasks, each contributing to an enhanced comprehension of emotional expressions and their computational detection. The scope of these tasks encompasses the recognition of emotional expressions [22, 24], the detection of facial action units [24, 25], regression analysis of valence and arousal [23, 24, 26, 45], and, more recently, the estimation of emotional mimicry [7]. These tasks are systematically organised within the Affective Behavior Analysis in-the-wild (ABAW) Challenge [19–21, 27], which, in its latest iteration [28], has introduced emotional mimicry in a more nuanced way as a novel area of exploration, reflecting the field's evolving focus and expanding methodologies.

Grósz et al. [14] present an multimodal approach on the MuSe Mimic dataset by integrating audio, visual, and textual data. They utilised fine-tuned Wav2vec 2.0 features to analyse audio inputs, extracted facial action units from video data to interpret visual cues, and employed an ELECTRA text encoder [8] for textual feature extraction. These diverse data streams were then synergistically combined using a late fusion technique.

In contrast to a conventional late fusion method, Ding et al. [11] implement a specialised fusion block with cross-attention mechanisms for the integration of multimodal data. This novel component is specifically trained to manage the fusion of different modalities, enhancing the model's capability to identify key features across the varied data types. This method represents a more nuanced approach to feature integration, aiming to improve the overall efficacy.

Adopting a multimodal transformer with cross-attention for data fusion, Guofeng et al. [43] distinguishes itself by utilising features from a large language model rather than conventional text encoders. This approach provides a more nuanced understanding of text, enriching the model's interpretative depth within a multimodal context.

While prior research often utilised the standard Wav2Vec 2.0 model, sometimes without any fine-tuning, Chen et al. [35] highlighted the necessity for task-specific adjustments. They demonstrated that the Wav2Vec 2.0 model, originally trained for automatic speech recognition (ASR), requires fine-tuning to better align with the nuances of emotion recognition tasks, due to the distinct nature of these applications. Similarly, Pepino et al. [6] explored the integration of additional audio features, such as eGeMAPS [13], with the Wav2Vec 2.0 framework. This combination offers a structured approach to enhance the model's performance by providing supplementary contextual cues for emotion analysis.

Building upon similar research, Wagner et al. [40, 41] conducted a comprehensive analysis across various tasks and models, revealing that extra fine-tuning aimed at automatic speech recognition (ASR) fails to enhance performance. This finding implies that the default Wav2Vec 2.0 model might not be optimally configured for emotion recognition tasks. Accordingly, their study highlights the importance of domain-specific training to better prepare the model for emotion detection.

Recent advancements in affective computing reveal that audio features are pivotal for recognising emotions, with multimodal integration highlighting the complexity and computational demands of current models. The use of advanced fusion techniques to combine diverse data underscores the challenge of balancing sophisticated analysis with the high computational needs of these complex architectures.

3. Dataset and Challenge

3.1. Challenge

The dataset [28] for the Emotional Mimicry Intensity Challenge (EMI-Challenge) includes over 30 hours of audiovisual content from 557 participants, recorded in natural settings using webcams. Participants were prompted to mimick seed videos showing a person expressing a particular emotion. After that, they had to rate the intensity of the resulting emotional experience on a 0-100 scale. Participants evaluated videos displaying emotional mimicry without knowing the intended emotion to be mimicked, ensuring their judgements remained unbiased. The dataset is split into training (8072 videos, approx. 15 hours), validation (4588 videos, approx. 9 hours), and test set (4586 videos, approx. 9 hours) without speaker overlap. Training and validation sets come with annotations, while test set predictions are submitted for evaluation. The dataset includes

detected faces at 6 fps, features from Vision Transformer (ViT) [3, 12] for video and Wav2Vec 2.0 features [1] and the raw videos and audios. Six different emotional expressions are annotated, namely: “Admiration”, “Amusement”, “Determination”, “Empathic Pain”, “Excitement”, and “Joy”.

The performance on the respective data-split (ρ_{VAL}, ρ_{TEST}) as reported in Table 1 and Table 2 in this task is measured with the Pearson’s Correlation Coefficient $\rho \in [-1, 1]$ averaged over all predicted emotions:

$$\rho = \frac{1}{6}\sum_{i=1}^{6} \rho_i \,, \tag{1}$$

with ρ_i as

$$\rho_i = \frac{Cov(X_{i,pred}, Y_{i,label})}{\sqrt{Var(X_{i,pred})}\sqrt{Var(Y_{i,label})}} \,. \tag{2}$$

In the boxplots depicted in Figure 1 for both the training and validation datasets, we observe a significant imbalance in the distribution of regression targets. The distribution for validation also mainly differs from training in the classes “Determination” and “Joy”.

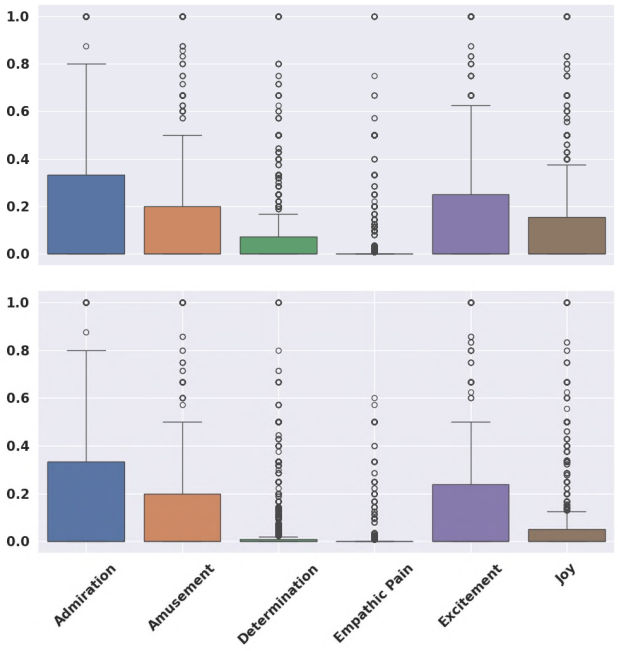

Figure 1. **Boxplots of the label distribution:** train (upper plot) and validation (lower plot).

3.2. Descriptive Data Analysis

The data in both plots are heavily skewed, with a vast majority of values congregating near zero. This skewness highlights the relative rarity of regression targets assigned the value ’1’, setting them apart as the less frequent outcomes

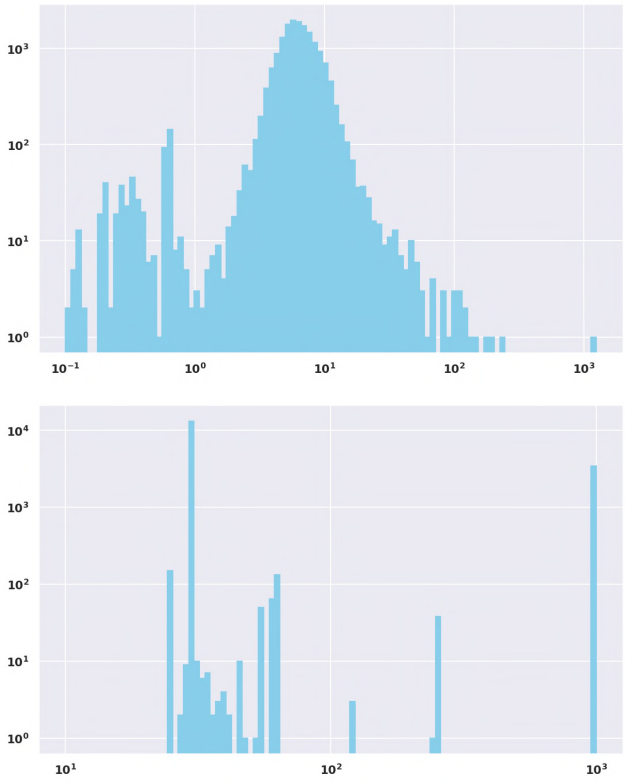

Figure 2. **Video duration’s in seconds (top) and fps (bottom) distribution over all data splits in log-log scale.**

amidst a dominant backdrop of values near zero - accounting for 55.3% to 89.9% of labels for training, respectively 34.1% to 52.4% for validation. Such a distribution presents challenges for learning algorithms, particularly in terms of accurately predicting or learning from these rare extreme values.

Audios are of constant quality - mono, $16kHz$ sampling rate, and $64kb/s$ bitrate.

Figure 2 highlights a key challenge in video analysis: the variability in video lengths and recording diversity, which impacts the stability of frames per second (fps) and significantly increases memory demands for batch processing videos of different lengths, or even for videos of same length with highly different fps. This variability introduces complexities for deep learning models, which rely on consistent input data formats and sizes for optimal training and inference.

For the main part of video lengths, we see a Gaussian distribution, with the left tail standing out. The distribution ranges from $0.1s$ to $1249.9s$ with $6.93s$ on average. 533 (3.1%) are shorter than $1s$, 753 (4.4%) longer than $12s$. Former are likely an issue of data quality, as mimicking an emotion in less than $1s$ seems infeasible, or rather is the result of recording errors. The latter have to be considered

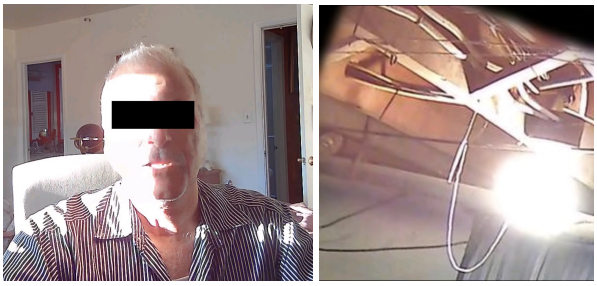

Figure 3. **Example of particular challenging videos:** overall quality, *e.g*. illumination (left, manual crop), affects downstream tasks, *e.g*. face detection (right).

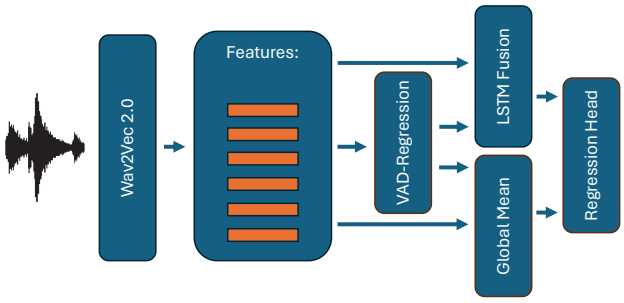

Figure 4. **Architecture overview of our approach.** We use a pre-trained Wav2Vec 2.0 model [40] with a Valence-Arousal-Dominance (VAD) module and extract the features as well as the VAD predictions. To leverage global context we use a global vector and fuse the temporal features in an LSTM.

for a trade-off between batch processing, finite VRAM, and seeing the whole length of the videos. We can safely say, that the video with a duration longer than 2 hours is either a recording error or mimics the same emotion for a very long time. The remaining 752 long videos have to be considered for the trade-off – we cut videos to a length of maximum $12s$, thereby seeing 95.6% of videos wholly, and the cut data amounts to 1.4% of total data – a good compromise.

The fps distribution is unexpected - since the recordings happen in a domestic webcam-based scenario, you would expect to see 2 to 3 bars – $25, 30$ and maybe 60 fps. The values range from 25 to 1000 fps, with an average of 226. 62.6% of videos are recorded at 30 fps, 20.1% at 1000 fps. There are also exotic values present like 29.83 or 62.5 fps. This means, that frames of different videos resembles a highly variable different amount of time, which affects audio alignment to said frames. Extracting faces at a fixed fps does not alleviate this completely, as you have to extract at a higher rate than the target fps, since the face detector can fail, thereby retaining some of the variance of time resembled per frame. *E.g*. for video "11437", originally recorded at 30 fps and a duration of $6.54s$, you expect $6.54s \cdot 6fps = 39$ frames, but are given 42 frames, resembling a frame rate of $42/6.54 = 6.42$. Analogously, for "15709" this results in $5.84fps$.

Figure 3 presents one of those issues where crucial facial features are occluded, adding another layer of complexity, failing downstream tasks. For deep learning algorithms, particularly those focused on facial recognition or emotion detection, occlusions can severely hinder the model's ability to accurately identify and analyse key features. This obstruction challenges the model's learning process as well as the above aligning frames to audio issue, which is why we focus on an audio-only approach.

4. Methodology

In our methodology, code publicly available here[1], we harness the capabilities of the pre-trained Wav2Vec 2.0 model [40, 41] as a core feature extractor. This choice is underpinned by Wav2Vec 2.0's proven efficiency in distilling important acoustic features, particularly for emotional speech analysis, a strength validated by numerous studies [11, 14, 43]. Our decision to rely solely on audio data is informed by the practical challenges encountered in therapeutic and real-world scenarios, such as the potential occlusion of facial expressions, which can compromise the reliability of visual cues. Moreover, while incorporating multimodal data can enrich the analysis, it invariably escalates computational demands.

To augment the model's ability to predict emotional mimicry intensity, which includes nuanced expressions like "Admiration", "Amusement", "Determination", "Empathic Pain", "Excitement", and "Joy", we employ a unimodal multi-task fusion approach by incorporating a Valence-Arousal-Dominance (VAD) prediction module, as shown in Figure 4. The inclusion of VAD is strategic, aimed at embedding an additional layer of emotional granularity. Valence captures the positivity or negativity of an emotion, arousal reflects the intensity of emotional activation, and dominance denotes the control level within the emotional experience. By integrating these dimensions, we aim to encapsulate a more comprehensive emotional spectrum, thereby providing a richer contextual basis for predicting specific emotional mimicry intensities.

End-to-end Description The raw mono-channel audios, sampled at $16kHz$, are cut to a maximum length of $12s$, *cf*. Section 3.2, and prepared by a Wav2Vec2.0-preprocessor. These preprocessed, variable length audios

[1] https://github.com/Skyy93/CVPR2024_abaw

are then batch-processed by a Wav2Vec2.0 model, fine-tuned to VAD regression [41], having continuous VAD values ranging from 0 to 1. The sequence lengths are reduced by the model to maximum of 599, for an initial maximum length of $12s \cdot 16kHz = 192000$. Both the 1024-dimensional model features, as well as the 3-dimensional VAD regressions, are fed concatenated as 1027 dimensions into mean pooling and an LSTM. Both eliminate the sequence dimension, with the former outputting a per audio global mean, and the latter yielding temporal aware features. These $2 \cdot 1027 = 2054$ features are then concatenated and finally processed by a regression head, consisting of a dense layer with tanh activation, and a projection layer, outputting the final 6 dimensional emotional mimicry intensities. For the architectural choice of our regression head we follow the VAD module from [41].

Implementation Details In our method we unfreeze the Wav2Vec 2.0 model together with the pre-trained VAD module. To improve the generalisation of the LSTM a dropout of 0.1 serves as additional regularisation. We adopt a learning rate of 1e-4, applying cosine decay. For our loss function, we use Mean Squared Error (MSE) because initial experiments with Concordance Correlation Coefficient (CCC) loss and Pearson Correlation Loss did not yield performance improvements. The training process spans 30 epochs, incorporating early stopping with a batch size of 32.

5. Evaluation

As demonstrated in Table 1, the proposed methodology surpasses the established baseline and secures the position of runner-up in the 6th Workshop and Competition on Affective Behavior Analysis in-the-wild. The baseline consists of pre-extracted Wav2Vec 2.0 features with a linear layer for the auditory modality. The vision features are extracted with a ViT and processed by a 3-layer gated recurrent unit (GRU) network. For the multimodal fusion the predictions were averaged. A detailed analysis of the architectural decisions is given in Section 6.1. Initially, the study explored the use of vision input as a foundational benchmark. However, due to limitations and challenges mentioned in Section 3.2 and Section 6.2, this approach is subsequently discontinued. Notably, incorporating vision input was found to detrimentally affect the performance of our unimodal model, particularly when subjected to joint fine-tuning.

For the test set, additional regularisation is applied through the incorporation of a 10% Dropout in the LSTM layer. The highest scores are achieved by training the model on a combined dataset of training and validation sets, thereby enhancing the model's capacity to address underrepresented emotions.

While our approach does not leverage the vision data, Savchenko et al. [38] use a variety of different Convolutional Neural Networks (CNNs) to encode the facial information of the dataset. The facial descriptor models are note fine-tuned and a simple linear layer is optimised on the statistical feature descriptor vector of the whole video. Wav2Vec 2.0 is also used for audio processing in their approach, albeit without fine-tuning or additional pre-training.

Yu et al. [44] employ vision features extracted by a ResNet-18, pre-trained on the AffectNet database [32], in conjunction with predicted Action Units and Wav2Vec 2.0 features. Considering the temporal characteristics of the video data, a Temporal Convolutional Network (TCN) is utilised for additional feature refinement across both modalities. To address long-range dependencies in the vision data, a Transformer encoder is applied to the vision features processed by the TCN. Consistent with the above research, it is observed that vision features underperform in comparison to audio features. To fuse the predictions of both modalities a late fusion strategy is employed, which provide additional performance gain, unlike our work.

A new feature extractor specifically designed for facial feature extraction was introduced by Zhang et al. [47], utilising a Masked Autoencoder [15] (MAE) trained over 800 epochs on an extensive composite dataset. This dataset, integrates multiple sources - Affect-Net [32], CASIA-WebFace [42], CelebA [30], IMDB-WIKI [37], and WebFace260M [48] - resulting in an expansive collection of 262 million images. Beyond facial analysis, their approach integrates audio encoding through VGGish [5] and textual feature extraction. The methodology utilises Transformer encoders tailored to each data modality, merging the outcomes via a voting strategy. This ensemble technique results in superior performance, albeit with the trade-off of substantial initial training requirements.

Across the compared approaches, a common finding is the underperformance of vision encoders compared to audio modalities, with sound consistently emerging as the most indicative modality for emotional analysis. Our methodology, distinct in its exclusive reliance on audio without the integration of vision or textual data, offers an efficient and streamlined approach for inferring the intensity of emotional mimicry, standing out for its simplicity.

6. Ablation

Our study adopts a unimodal approach, centering on the auditory modality as the primary source of input. This decision is underscored by the employment of the Wav2Vec 2.0 framework, renowned for its ability to capture rich, nuanced acoustic features from raw audio data. In this ablation we feature multiple experiments over our architectural choice and the challenges that arise when using a multimodal approach.

Model	Vision	Audio	ρ_{VAL}	ρ_{TEST}
Baseline [28]	X	-	-	.090
Baseline [28]	-	X	-	.240
Baseline [28]	X	X	-	.250
Savchenko et al. [38]	X	X	.289	.331
Yu et al. [44]	X	X	.328	.359
Zhang et al. [47]	X	X	**.463**	**.718**
$\text{Ours}_{\text{train}}$	X	-	.013	-
$\text{Ours}_{\text{train}}$	X	X	.198	-
$\text{Ours}_{\text{train}}$ (freezed)	-	X	.262	-
$\text{Ours}_{\text{train}}$	-	X	.386	.461
$\text{Ours}_{\text{train}}$ (w/ Dropout)	-	X	.389	.465
$\text{Ours}_{\text{train+val}}$	-	X	-	.522
$\text{Ours}_{\text{train+val}}$ (longer Training)	-	X	-	.554

Table 1. **Quantitative comparison of our approach.** We compare our solution with the Top-3 other solutions and achieve second place in the EMI Challenge.

6.1. Architectural Choices

In our study, we systematically evaluated various configurations of the Wav2Vec 2.0-large model. Our results, detailed in Table 2, demonstrate the incremental impact of different architectural elements on model performance, measured by the correlation coefficient (ρ_{VAL}).

The foundational experiment utilised the fine-tuning of a Wav2Vec 2.0-large model [1] and its multilingual derivate [9], yielding a ρ_{VAL} of 0.017 and 0.021. This initial outcome highlighted the limitations of using models pre-trained on standard Automatic Speech Recognition (ASR) tasks, such as those involving the LibriSpeech dataset [34], for complex emotional recognition tasks.

Subsequent experiments involved incrementally adding components to the model architecture, such as Global Vector, Regression Head, and LSTM layers. The introduction of a Regression Head and Global Vector improved the ρ_{VAL} to 0.356, indicating the significance of incorporating model components that enhance the representation of speech features relevant to emotion recognition.

Further incorporation of LSTM layers, known for their effectiveness in capturing temporal dependencies in data, resulted in a ρ_{VAL} of 0.375. A notable point from our findings is that neither the global vector nor the VAD (Valence, Arousal, Dominance) module significantly enhances performance on their own. However, when integrated, they collectively offer an additional enhancement.

The most comprehensive model configuration, which included all evaluated components (Global Vector, Regression Head, LSTM, and VAD Head), achieved the highest ρ_{VAL} of 0.386.

Our experimental results underscore the critical role of tailored pre-training and architectural design in enhancing model performance on specialised tasks like estimating the emotion mimicry intensity. It is evident that generic ASR pre-training is insufficient for complex tasks requiring nuanced emotional understanding, thereby highlighting the need for domain-specific adaptations in model training and architecture.

Model	Global Vector	Reg. Head	LSTM	VAD Head	ρ_{VAL}
W2V2-L [1]	X	-	-	-	.017
W2V2-L XLSR [9]	X	-	-	-	.021
W2V2-L Audeering [40]	X	-	-	-	.342
W2V2-L Audeering [40]	X	X	-	-	.356
W2V2-L Audeering [40]	-	X	X	-	.375
W2V2-L Audeering [40]	X	X	X	-	.377
W2V2-L Audeering [40]	-	X	X	X	.375
W2V2-L Audeering [40]	X	X	X	X	.386

Table 2. **Comparison of our design choices.**

6.2. Exploratory Findings

Zero-padding, *i.e.* filling missing video data with black images or audio with silence to create fixed length batch tensors, was not only detrimental to memory efficiency, thereby batch size and training duration, but also to training itself – the padding amounted to 34% of batch data on average, often disallowing the model to learn anything at all. ρ in the region of -0.1 was not uncommon.

Correlation as loss instead of MSE, *i.e.*

$$L = 1 - \rho\,,\; L \in [0, 2]\,, \tag{3}$$

did not work aswell. With the huge imbalance in the labels towards a single value, *cf*. Section 3.2, the models quickly learn to make a constant prediction, thereby having an undefined variance (*cf*. Equation (2)), therefore having an undefined loss, blocking training via backpropagation completely.

Fine-tuning a generally pretrained vision model, *e.g.* Convnext [31] or DinoV2 [33], yielded some improvement for ρ from -0.10 to -0.01, but remained underwhelming to a degree, that we doubted the soundness of our pipeline. To check, we trained and validated using a 2:1 split both on official training and official validation split – in both cases ρ rose to about 0.18. This lead us to the assumption, that the pipeline itself works, but that the vision models do not generalise at all from training to validation split, but immediately overfit to the data.

To mitigate the variance of the time between face detection frames, *cf*. Section 3.2, we tried face detection using BlazeFace [2] at a constant framerate. Also increasing the crop area from 160x160 to 256x256, and aligning the detected faces afterwards. But this did not change the performance of beforehand vision models noticeably.

We then replaced the vision models with models trained for facial expression recognition, *e.g.* LibreFace [4] and EfficientNet [39]. The performance did somewhat improve to

our best vision-only performance of $\rho = 0.013$, respectively multimodal of $\rho = 0.198$, but they still disappoint.

This then lead us to dropping the vision modality completely, yielding a 32% boost to the performance ($\rho = 0.262$) when training on audio only.

7. Discussion

In our study, we concentrated on analysing emotional mimicry intensity through audio data, moving away from the common focus on facial expressions in affective computing. Despite having a comprehensive multimodal dataset, we found that adding facial images to the analysis decreased its effectiveness, as shown by lower Pearson correlation coefficients when including vision compared to audio-only results. Our findings emphasise the need to choose the right modality for emotional analysis and hint at audio's unique potential in this field. Notably, our approach stood out among challenge participants by employing a unimodal strategy, prioritising computational speed and resource efficiency. This distinctive choice underlines the potential of streamlined, audio-focused analysis in settings where efficiency is paramount. Additionally, an intriguing avenue for future research could be the development of multimodal models that not only process audio signals but also interpret the textual content within these signals, enriching the emotional analysis. Further research could explore aligning audio with facial expressions, addressing the challenging task of effectively integrating these modalities.

8. Conclusion

Our study utilised the pre-trained Wav2Vec 2.0 model for emotional speech analysis, focusing on audio data to address real-world challenges like facial occlusion. We enhanced emotion detection by integrating a Valence-Arousal-Dominance (VAD) module with our model, aiming for a deeper emotional understanding. Our multi-task approach, validated through an extensive ablation study, showed promising results in predicting emotional mimicry intensities. This work, which achieved second place in the Emotional Mimicry Challenge at the 6th Affective Behavior Analysis in-the-wild Workshop, highlights the potential of audio-focused analysis in affective computing.

Acknowledgement

The authors gratefully acknowledge the computing time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster Monacum One at the University of the Bundeswehr Munich.

This work was partially funded by the KodiLL project (FBM2020, Stiftung Innovation in der Hochschullehre).

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 3, 6

[2] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus, 2019. 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[4] Di Chang, Yufeng Yin, Zongjian Li, Minh Tran, and Mohammad Soleymani. Libreface: An open-source toolkit for deep facial expression analysis, 2023. 6

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 5

[6] Li-Wei Chen and Alexander Rudnicky. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[7] Lukas Christ, Shahin Amiriparian, Alice Baird, Alexander Kathan, Niklas Müller, Steffen Klug, Chris Gagne, Panagiotis Tzirakis, Lukas Stappen, Eva-Maria Meßner, et al. The muse 2023 multimodal sentiment analysis challenge: Mimicked emotions, cross-cultural humour, and personalisation. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, pages 1–10, 2023. 1, 2

[8] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020. 1, 2

[9] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020. 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[11] Chaoyue Ding, Daoming Zong, Baoxiang Li, Song Zhang, Xiaoxu Zhu, Guiping Zhong, and Dinghao Zhou. Multimodal sentiment analysis via efficient multimodal transformer and modality-aware adaptive training strategy. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, page 11–17, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2, 4

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[13] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2): 190–202, 2016. 2

[14] Tamás Grósz, Anja Virkkunen, Dejan Porjazovski, and Mikko Kurimo. Discovering relevant sub-spaces of bert, wav2vec 2.0, electra and vit embeddings for humor and mimicked emotion recognition with integrated gradients. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, page 27–34, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2, 4

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5

[16] Ursula Hess and Agneta Fischer. Emotional mimicry as social regulation. *Personality and social psychology review*, 17 (2):142–157, 2013. 1

[17] Ursula Hess and Agneta Fischer. Emotional mimicry as social regulator: theoretical considerations. *Cognition and Emotion*, 36(5):785–793, 2022. PMID: 35920780. 1

[18] Ursula Hess, Pierre Philippot, and Sylvie Blairy. Mimicry: Facts and fiction. *The social context of nonverbal behavior*, pages 213–241, 1999. 1

[19] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 2

[20] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023.

[21] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023. 2

[22] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 2

[23] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 2

[24] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 2

[25] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 2

[26] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 2

[27] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023. 2

[28] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024. 2, 6

[29] Katja U Likowski, Andreas Mühlberger, Beate Seibt, Paul Pauli, and Peter Weyers. Modulation of facial mimicry by attitudes. *Journal of experimental social psychology*, 44(4): 1065–1072, 2008. 1

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 6

[32] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 5

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 6

[34] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. 6

[35] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021. 2

[36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 1

[37] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018. 5

[38] Andrey V. Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction, 2024. 5, 6

[39] Andrey V. Savchenko, Lyudmila V. Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13 (4):2132–2143, 2022. 6

[40] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0, 2022. 2, 4, 6

[41] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 4, 5

[42] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5

[43] Guofeng Yi, Yuguang Yang, Yu Pan, Yuhang Cao, Jixun Yao, Xiang Lv, Cunhang Fan, Zhao Lv, Jianhua Tao, Shan Liang, and Heng Lu. Exploring the power of cross-contextual large language model in mimic emotion prediction. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, page 19–26, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2, 4

[44] Jun Yu, Wangyuan Zhu, and Jichao Zhu. Efficient feature extraction and late fusion strategy for audiovisual emotional mimicry intensity estimation, 2024. 5, 6

[45] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 2

[46] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 1

[47] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge, 2024. 5, 6

[48] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5