Beyond Static Emotions: Leveraging Multitask Learning to Model Dynamics of Dimensional Affect in Speech

Yuxuan Zhang[1,2], Hippolyte Fournier[1], Ruslan Kalitvianski[2], Marco Dinarelli[1], and Fabien Ringeval[1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
{yuxuan.zhang,hippolyte.fournier,marco.dinarelli,fabien.ringeval}@univ-grenoble-alpes.fr
[2] ELOQUANT, 5 allée de Palestine, 38610 Gières, France
ruslan.kalitvianski@eloquant.com

Abstract. Dimensional affect prediction from speech has traditionally relied on acoustic features to estimate continuous affect representations (e.g., arousal, valence) at each time step. However, affect evolves dynamically over time, and incorporating temporal information may improve prediction accuracy. This study investigates emotional dynamics in speech emotion recognition using multitask learning, where a model jointly predicts both the affect state and its temporal derivative. Experiments on the RECOLA and SEWA datasets show that incorporating dynamic information improves affect state prediction, particularly for valence, known to be challenging to model from audio alone. While CCC scores for affect dynamic predictions remain lower than those for affect state predictions, results indicate that learning dynamics as an auxiliary task enhances affect state estimation over time. These findings underscore the importance of modelling emotional dynamics to capture the temporal evolution of affect.

Keywords: affective computing, dimensional affect, computational paralinguistics

1 Introduction

In recent decades, a growing consensus in the literature has highlighted the importance of viewing affect as an episode rather than as a static state [29, 39, 34, 19, 4]. More specifically, this perspective emphasises the need to consider the dynamics of states within an affective episode in order to fully understand it. A clear illustration of this postulate is the difference between joy and relief: relief differs from joy in that a transient state of uncertainty precedes the fulfillment of a desired event [15]. In other words, these two emotions cannot be properly distinguished without considering the temporal dynamics of the episode. Applied to affect prediction, this principle necessitates capturing affect representations continuously rather than as isolated points in time.

Traditional approaches in speech emotion recognition (SER) primarily focus on mapping acoustic features to affect labels at each timestamp independently, overlooking the temporal evolution of emotions. However, emotional states are inherently dynamic and context-dependent [30], influenced by preceding and ongoing interactions.

Modelling affect dynamics in continuous speech is particularly challenging due to the temporal dependencies between affect states. Several studies have explored Recurrent Neural Networks (RNNs) and attention-based architectures to capture sequential dependencies in emotion trajectories [20, 23, 22]. Yet, most models still optimise for static predictions and fail to explicitly encode emotional changes over time, limiting their ability to model emotional transitions, escalation, or decay.

To address this gap, we propose a multitask learning (MTL) approach that jointly predicts affect and its temporal derivative, encouraging the model to learn both the current emotional state and its temporal variation. By explicitly integrating emotional dynamics, we hypothesise that MTL will enhance affect prediction, particularly for valence, which is known to be difficult to infer from speech alone [37].

Our contributions are as follows: (i) We first introduce a multitask learning framework for speech emotion recognition that simultaneously predicts dimensional affect and its temporal evolution. (ii) We evaluate our approach on two benchmark datasets, RECOLA [28] and SEWA [18], across three languages – French, German, and Hungarian –, and analyse how dynamic information impacts valence and arousal predictions. (iii) We demonstrate that explicitly modelling affect dynamics systematically improves affect prediction, achieving state-of-the-art results using simple GRU models, despite previous work employing more complex architectures. These findings highlight the importance of capturing the temporal flow of emotions.

The remainder of the paper is structured as follows: Section 2 reviews related work on continuous affect prediction and multitask learning. Section 3 presents our proposed approach, followed by the experimental setup and results in Section 4. Section 5 discusses findings and limitations and concludes with future research directions.

2 Related Work

Predicting affect continuously over time remains a longstanding challenge in affective computing [40]. Traditional dimensional affect prediction models map either handcrafted or self-supervised acoustic features to valence and arousal scores at each time step [1]. However, affect states tend to evolve gradually rather than change abruptly, making temporal modelling crucial for improving prediction accuracy.

Among affect dimensions, valence – representing intrinsic pleasantness – remains one of the most challenging to predict from speech alone [37]. Unlike arousal, which is closely tied to acoustic energy and prosody, valence is often conveyed through semantic and contextual cues. Prior research has shown that multimodal approaches, e.g., integrating facial expressions and lexical content, improve valence prediction [27, 31], but these approaches rely on data such as video that is not always available, for instance in call centre settings.

To address these difficulties, several studies have leveraged sequential deep learning models such as Long Short-Term Memory (LSTM) networks [12] and Gated Recurrent Units (GRUs) [5] to capture long-range dependencies in affect signals. For instance, an end-to-end deep learning approach using LSTMs has been proposed to model speech-based emotional trajectories [33]. Similarly, convolutional-recurrent hybrid models have been explored for end-to-end SER [36]. More recently, self-attention-

based transformers have been investigated for emotion sequence modelling [9, 32], showing promising results in capturing contextual dependencies over time.

Multitask learning (MTL) has been widely adopted in affective computing to improve emotion recognition by leveraging shared representations across related tasks. Prior work has explored MTL for jointly predicting multiple emotion dimensions, such as valence, arousal, and dominance [25]. Other studies have incorporated speaker identity, contextual information, or physiological signals as auxiliary tasks to improve generalisation [13].

A particularly relevant approach involves training models to predict both an affect state and its temporal evolution. The AVEC 2019 State-of-Mind sub-challenge proposed dynamic representations of affect to better capture emotional changes in personal storytelling [27]. Similarly, emotional dynamics have been explicitly modelled in conversational settings using multitask frameworks that jointly classify emotion labels, emotion shifts, and sentiment [38]. Other studies have leveraged context-aware networks to more effectively capture the evolution of emotions in conversations [11, 35].

Despite recent advances, few studies explicitly integrate affect dynamics into time-continuous emotion prediction. While fine-tuning pretrained transformer models on dimensional affect has been explored [37], these approaches often overlook its temporal evolution. A notable prior approach to addressing this limitation is the use of neural ordinary differential equations. Initially explored in related domains such as time-series forecasting [14], this method was later applied to continuous emotion recognition by reformulating the prediction task as an ordinary differential equation involving both the emotion signal and its derivative, thereby implicitly modelling affect dynamics [7]. Building on this direction, our work introduces a multitask learning framework that jointly predicts affect states and their temporal derivatives, enabling the model to explicitly learn how emotions evolve over time. We hypothesise that this approach improves continuous affect state prediction, even in unimodal speech settings.

3 Methodology

This section details the preprocessing of affect annotations and training features, followed by the architecture and objectives of our MTL framework. Figure 1 illustrates the overall training methodology. We first obtain affect dynamics references by computing the temporal derivatives of the Gold Standard affect annotations. The models are then trained to jointly predict dimensional affect and its temporal evolution, leveraging both handcrafted and self-supervised features.

3.1 Temporal Differentiation of Dimensional Affect

In order to model affect dynamics from the Gold Standard annotation, we approximate the temporal derivative of each affect dimension E by computing its discrete-time differential for each time step i:

$$\left(\frac{dE}{dt}\right)[i] \approx \frac{E[i] - E[i-1]}{t[i] - t[i-1]} \tag{1}$$

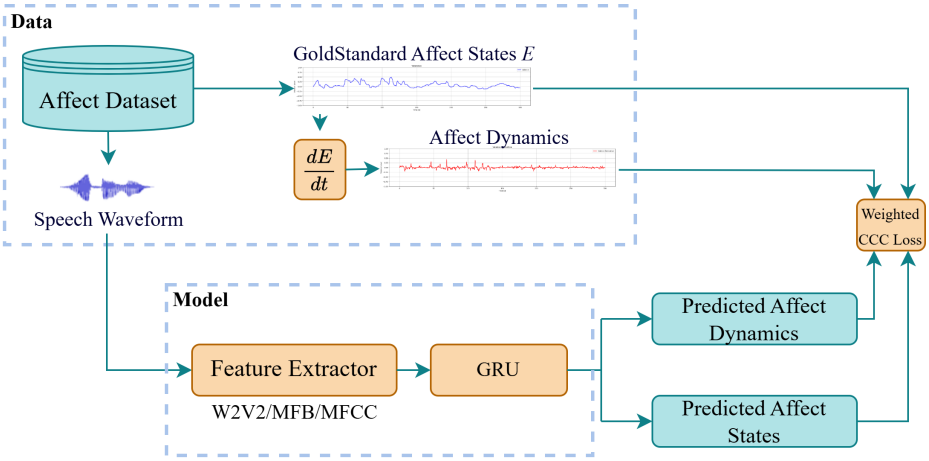


Fig. 1: The proposed MTL training methodology applied on a single affect dimension. Affect dynamics are derived by computing the temporal derivative of the Gold Standard affect annotations. Models are then trained to jointly predict both continuous affect states and their dynamics.

3.2 Audio Features

Handcrafted Features Traditional speech-based affect recognition relies on handcrafted low-level descriptors (LLDs), including spectral, cepstral, and prosodic features extracted over sliding windows. We use 40 Mel-Filter Bank (MFB) features, standardised to zero mean and unit variance. Additionally, we extract MFCCs (1–13) with their first and second-order derivatives. Following AVEC 2019 [27], mean and standard deviation are computed over 4-second windows with a 100 ms hop size. To align with affect annotations (4 Hz), all features are down-sampled via averaging over a sliding window, with parameters set according to the down-sampling rate.

Self-Supervised Representation Learning from Speech Recent studies have demonstrated that self-supervised speech models, such as Wav2Vec2 (W2V2) [3], provide robust representations for affect recognition, outperforming traditional handcrafted features [10, 37]. In our study, we extract W2V2 embeddings from the audio input as features. For multilingual settings, we use language-specific pretrained models, if available, or multilingual XLSR models [6] otherwise. These models are used purely as feature extractors, with no fine-tuning for speech emotion recognition. As with handcrafted features, extracted speech representations are down-sampled by averaging over a sliding window to match the affect annotation rate (4 Hz).

3.3 Model Design

Our MTL model consists of a 2-layer GRU with a hidden size H, followed by an output head that predicts both affect states and their temporal derivatives. Instead of using N separate output heads (each of size 1), we opt for a single shared output head to

encourage cross-task knowledge sharing; the output head is a single linear layer with a $tanh$ activation function, where the input size is H and the output size is N.

We implement two model configurations: (i) Mono-dimensional model: Handles a single affect dimension ($N = 2$), predicting both affect and its dynamics. (ii) Multi-dimensional model: Handles D affect dimensions predicting both affect states and their derivatives within a single model ($N = 2 \times D$). For the mono-dimensional models, we experiment with hidden sizes $H = \{16, 32\}$. For the multi-dimensional models, we also experiment with $H = 64$ to accommodate the increased number of tasks.

3.4 Loss Function and Metric

We optimise and evaluate our model using the Concordance Correlation Coefficient (CCC) [21], which measures agreement between the prediction y and the gold standard x:

$$\rho_{ccc} = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{2}$$

where μ_x and μ_y are the means, σ_x and σ_y the variances, and σ_{xy} is the covariance. The CCC loss is defined as:

$$L_{ccc} = 1 - \rho_{ccc} \tag{3}$$

We train an MTL model to jointly predict both affect states and their derivatives (dynamic) using the CCC loss. To balance these tasks, we introduce a weight parameter $w \in [0, 1]$ defining the mono-dimensional loss L_{mono} as:

$$L_{mono} = (1 - w)L_{state} + wL_{dynamics} \tag{4}$$

where L_{state} is the CCC loss computed between the model's prediction $\hat{E}$ and the gold standard affect state E, and $L_{dynamics}$ is the CCC loss computed between the predicted affect derivative $\frac{d\hat{E}}{dt}$ and the gold standard affect derivative $\frac{dE}{dt}$.

For multi-dimensional setups, the loss is averaged over D affect dimensions:

$$L_{multi} = \frac{1}{D}\sum_{d=1}^{D} L_{mono_d} \tag{5}$$

4 Experiments and Results

4.1 Datasets

We evaluate our approach on two well-established benchmark datasets for dimensional affect prediction – RECOLA and SEWA –, which provide a controlled vs. in-the-wild comparison for affect prediction across three languages: French, German, and Hungarian.

The REmote COLlaborative and Affective interactions (RECOLA) dataset [28] is a multimodal corpus of 3.8 hours of spontaneous, collaborative interactions in French,

collected from 46 participants engaged in dyadic problem-solving tasks. Recorded in a controlled, noise-free environment with a high-quality microphone setup, RECOLA includes valence and arousal annotations, rated continuously over time by multiple annotators. We use the training, development, and test splits from [33].

The SEWA corpus [18] contains over 33 hours of audio-visual recordings from 398 participants engaged in spontaneous discussions about advertisements they had watched, providing naturalistic affective expressions in a conversational setting. Unlike RECOLA, SEWA was recorded in the wild, with speech captured in varied acoustic environments using different recording devices, such as laptop and smartphone microphones, making it more representative of real-world affective speech. The dataset spans six cultural groups, each with approximately 66 participants, and includes valence, arousal, and liking annotations, rated continuously over time, similarly to RECOLA. We focus on the German (SEWA DE) and Hungarian (SEWA HU) subsets, following the same training, development, and test partitioning as in AVEC 2019 [27].

4.2 Experimental Setup

For each dataset, we extract MFB, MFCC, and W2V2 features. For French, we use W2V2 models pretrained on 2.9 k hours of French speech, available in two architectures: base (768-dimensional vectors) and large (1024-dimensional vectors) [10][3]. For German[4] and Hungarian[5], we use XLSR models fine-tuned on the respective languages.

The aim of our experiment is to assess the impact of weight parameter w in the loss function L_{mono} as defined in Equation 4, on the model's ability to predict the affect state and its dynamic for arousal and valence. For each model configuration (cf. section 3.3) and for each feature set (cf. section 3.2), we train models using values of w within the range $[0, 1]$ in increments of 0.1: Setting $w = 0$ corresponds to training solely on an affect state, while $w = 1$ trains exclusively on its dynamics. Increasing w gradually shifts the model's focus toward optimising affect dynamics prediction.

All training uses a unit batch size with the Adam optimiser [17] and a learning rate of 6e-4. Training runs for a maximum of 250 epochs, with early stopping after 30 epochs of no improvement on the development set. These values were selected through hyper-parameter tuning with $w = 0.5$.

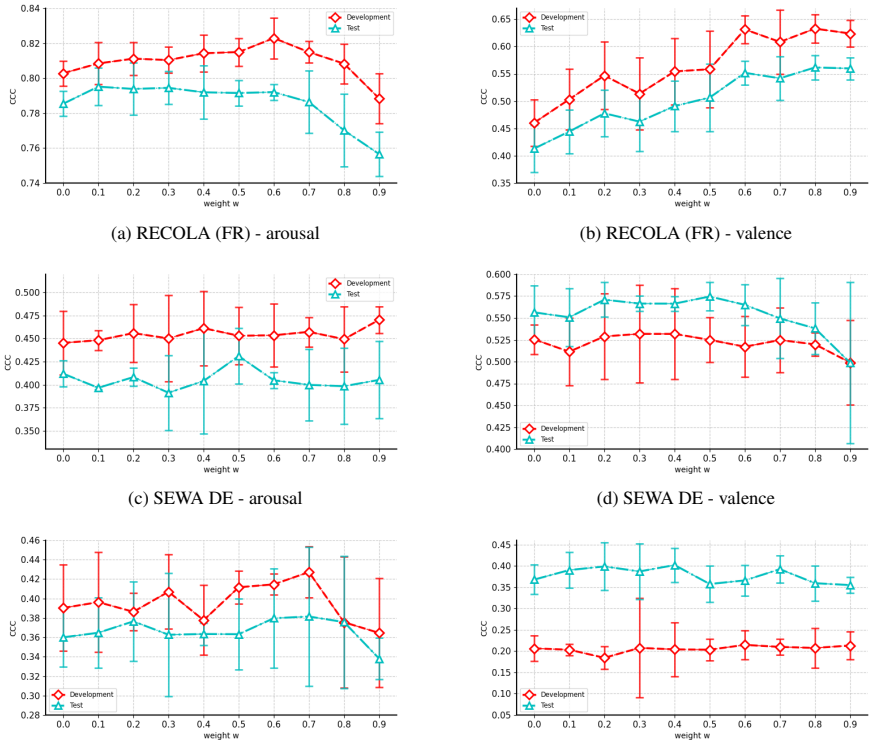
4.3 Experimental Results

Table 1 reports the CCC scores for predicting affect (left) and its dynamics (right) on the test set, using the value of w that yielded the best performance on the development set. Results are provided for both mono-dimensional and multi-dimensional setups, across all features and corpora, and are compared to models trained exclusively on affect states ($w = 0$). Paired permutation tests with 1000 permutations between the best model and

[3] https://huggingface.co/LeBenchmark

[4] https://huggingface.co/facebook/wav2vec2-large-xlsr-53-german

[5] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-hungarian



(a) RECOLA (FR) - arousal

(b) RECOLA (FR) - valence

(c) SEWA DE - arousal

(d) SEWA DE - valence

(e) SEWA HU - arousal

(f) SEWA HU - valence

Fig. 2: Evolution of the CCC score, averaged over multi-dimensional models with W2V2 features, as a function of weight w for predicting arousal and valence on the development and test partitions of the RECOLA (FR), SEWA DE, and SEWA HU corpora, with error bars indicating the 95% confidence interval. Scores for $w = 1$ are not reported for visibility purposes as the values are close to zero.

the model trained solely on affect states are performed to further evaluate the significance of the improvement achieved by the new approach. A p-value less than .05 indicates significance and is marked with an asterisk (*) next to the corresponding CCC. Figure 2 illustrates the evolution of average CCC scores for multi-dimensional models with W2V2 features, on both the development and test sets, as a function of w.

In most cases, the best performance on the development set is achieved when affect dynamics are included ($w \neq 0$), particularly for valence. However, determining the optimal value of w remains challenging, as different corpora, languages, and affect dimensions exhibit varying sensitivity to this parameter. On RECOLA (FR), training models to predict both affect states and their dynamics consistently improves valence prediction. In contrast, the benefits are less evident for arousal, as well as for SEWA DE and SEWA HU. These trends are also reflected in the evolution of average CCC scores shown in Figure 2. Additionally, it is important to note that the value of w yielding the

Table 1: CCC scores for predicting arousal and valence, as well as their dynamics, using different acoustic features (MFCC, MFB, W2V2) on the test sets of RECOLA (FR), SEWA DE, and SEWA HU. Left: predictions evaluated against affect states; right: against their temporal derivatives. Rows labeled $w = 0$ show models trained only on affect states; rows labeled best w indicate optimal weights based on development set. Best results per affect dimension are in bold. For affect state predictions, significant improvements over $w = 0$ ($p < .05$) are marked with *.

		Affect State						**Affect Dynamics**					
		Mono-dimensional			**Multi-dimensional**			**Mono-dimensional**			**Multi-dimensional**		
		MFCC	**MFB**	**W2V2**	**MFCC**	**MFB**	**W2V2**	**MFCC**	**MFB**	**W2V2**	**MFCC**	**MFB**	**W2V2**
		RECOLA (FR)											
Arousal	CCC best w	.799	.799	.795	.784	.799	.791	.470	.615	**.629**	.471	.620	.616
	CCC $w = 0$	.760	.799	**.807**	.768	.799	.787	.061	.350	.004	.053	.168	.124
	best w	0.8	0.1	0.9	0.5	0.0	0.6	1.0	1.0	1.0	1.0	0.7	0.8
Valence	CCC best w	.471	.485	.577*	.492	.428	**.578***	.260	.450	.559	.289	.426	**.562**
	CCC $w = 0$	.416	.440	.427	.446	.395	.426	.000	.071	.002	-.002	.058	.044
	best w	0.4	0.9	0.5	0.4	0.3	0.6	0.4	0.9	1.0	0.8	0.7	0.8
		SEWA DE											
Arousal	CCC best w	.292	.341	.377	.218	.285	**.440**	.070	.145	.262	.046	.131	**.267**
	CCC $w = 0$	.165	.354	.406	.336	.285	.401	-.004	-.004	.000	.001	.019	.012
	best w	0.6	0.2	0.9	0.8	0.0	0.9	0.8	0.8	0.3	0.8	0.8	1.0
Valence	CCC best w	.190	.304	.561	.261	.271	**.574**	.081	.119	.352	.0.095	.165	**.387**
	CCC $w = 0$	.190	.396	.548	.388	.271	.549	.001	-.002	.000	.001	.049	.001
	best w	0.0	0.5	0.8	0.8	0.0	0.3	0.8	0.5	1.0	0.8	0.8	1.0
		SEWA HU											
Arousal	CCC best w	.197	.241	.386	.092	.028	**.441***	.048	.122	.170	.023	.137	**.177**
	CCC $w = 0$	.126	.241	.379	.092	.121	.375	-.011	.001	-.004	.001	-.009	-.002
	best w	0.8	0.0	0.5	0.8	0.8	0.7	0.8	0.8	0.8	0.7	0.8	0.8
Valence	CCC best w	.152	.192	.345	-.126	.081	**.394**	.038	.050	.121	.045	.081	**.137**
	CCC $w = 0$	.031	.145	.332	.026	.107	.386	.007	.004	.004	.000	.006	.003
	best w	0.4	0.7	0.1	0.7	0.4	0.8	0.8	0.2	0.8	0.6	0.3	0.8

best performance on the development set does not always generalise optimally to the test set. This helps explain cases where the best development performance is achieved with $w \neq 0$, yet the CCC on the test set remains lower than that of the model trained solely on affect states.

Comparing mono-dimensional and multi-dimensional models, we observe that the best performance in predicting affect states (highlighted in bold in Table 1) is mostly achieved when jointly training on arousal, valence, and their dynamics. This suggests that the four tasks handled by the multi-dimensional model are correlated, facilitating knowledge sharing between tasks in the hidden representations, which benefits valence prediction—known to be challenging to model from speech alone. This effect is particularly noticeable for valence prediction on the RECOLA (FR) dataset, and is less evident on SEWA DE and SEWA HU, where the benefits of multi-dimensional models are primarily observed when trained on W2V2 features.

The observations above demonstrate that feature choice plays a crucial role. Handcrafted features (MFCC, MFB) appear to provide less informative representations as the number of tasks—that is, the complexity of the prediction—increases, resulting in little to no improvement over traditional methods or the mono-dimensional approach when

jointly training on multiple dimensions. These findings highlight the advantages of self-supervised pretrained models in SER. Although not fine-tuned for this task, W2V2 embeddings consistently outperform handcrafted features, especially in multi-dimensional models.

As seen on the right side of Table 1, CCC scores for affect dynamics prediction remain lower than those for affect state prediction, even in the best cases, which mostly occur when w is closer to 1.0, i.e., when training is more biased toward affect dynamics. Moreover, training exclusively on affect states ($w = 0.0$) leads to poor CCC scores when tested on affect dynamics, and *vice versa*. Interestingly, this contrasts with findings from AVEC 2019 [27], where models trained on dynamics performed better when tested on affect states in a non-time-continuous setting.

Finally, we compare our approach—specifically the models using W2V2 features and multi-dimensional architectures—with state-of-the-art results reported in Table 2. It is important to note that some RECOLA results were obtained using the 18-subject version (9 for training, 9 for validation) [26, 7], whereas our models are trained and evaluated on the full 46-subject set, reducing the risk of overfitting while still achieving state-of-the-art performance. We also note that our method significantly improves CCC scores for valence prediction across all three datasets compared to previous works.

Table 2: CCC of proposed (W2V2 + multi-dimensional model) and state-of-the-art methods.

Dataset	Method, Features + Model	Arousal	Valence
RECOLA (FR)	Praveen et al. 2022 [26], DFT + CNN	**.822**	.463
	Dang et al. 2024 [7], BoAW-MFCC + CD-NODE	.782	.506
	Parcollet et al. 2024 [24], W2V2 + GRU	.664	.466
	Ours, W2V2 + MTL-GRU (best w)	.791	**.578**
SEWA DE	Ringeval et al. 2019 [27], BoAW-eGeMAPS + LSTM	.276	.325
	Ringeval et al. 2019 [27], MFCC + LSTM	.296	.288
	Ours, W2V2 + MTL-GRU (best w)	**.440**	**.574**
SEWA HU	Ringeval et al. 2019 [27], BoAW-eGeMAPS + LSTM	.250	.151
	Ringeval et al. 2019 [27], MFCC + LSTM	.159	-.019
	Ours, W2V2 + MTL-GRU (best w)	**.441**	**.394**

5 Discussion and Conclusion

Our study demonstrates that modelling emotion dynamics through multitask learning generally enhances the prediction of continuous dimensional emotions, particularly when combined with self-supervised pretrained speech representations. However, the optimal value of w on the development set does not always generalise well to the test set, highlighting the challenge of selecting this parameter. Moreover, the benefits of incorporating emotion dynamics are less pronounced for arousal, as well as for more complex in-the-wild datasets such as SEWA DE and SEWA HU. Noise introduced by the derivative operation makes training more challenging, particularly for the less controlled SEWA dataset, which may explain the limited improvement observed when

incorporating affect dynamics—unlike the more controlled RECOLA dataset. Future work should explore signal smoothing techniques to mitigate the negative impact of noisy affect dynamics, such as temporal averaging or convolutional layers with a sinc kernel acting as a low-pass filter with a trainable cut-off frequency [16, 2].

The multi-dimensional setup further improves performance by leveraging correlations between arousal, valence, and their respective dynamics, especially with W2V2 features. Our results confirm that pretrained self-supervised models consistently outperform handcrafted features, demonstrating their effectiveness even without fine-tuning for emotion recognition. Combined with our MTL framework, a simple GRU achieves state-of-the-art results on RECOLA and SEWA, despite past works using more powerful models such as Bi-LSTMs with access to future contexts. These findings highlight the potential of emotion dynamics modelling, multitask learning, and self-supervised pretrained models in advancing dimensional affect recognition from speech.

Despite these advancements, the choice of w remains both language- and affect-dependent, making it a crucial factor in model performance. Future work should explore adaptive strategies for dynamically tuning w and further investigate fine-tuning of W2V2 for SER. Moreover, it would be of interest to explore speech encoders that leverage semantic information, such as SONAR [8], as acoustic descriptors.

6 Acknowledgements

This work is part of a CIFRE thesis with the grant number 2023/1536 from the French National Association for Research and Technology (ANRT).

References

1. Alisamir, S., Ringeval, F.: On the Evolution of Speech Representations for Affective Computing: A brief history and critical overview. IEEE Signal Processing Magazine **38**(6), 12–21 (November 2021)
2. Alisamir, S., Ringeval, F., Portet, F.: Dynamic time-alignment of dimensional annotations of emotion using recurrent neural networks. arXiv e-prints p. 2209.10223 (2022)
3. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems. vol. 33, pp. 12449–12460 (2020)
4. Bledow, R., Rosing, K., Frese, M.: A dynamic perspective on affect and creativity. Academy of Management Journal **56**(2), 432–450 (2013)
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS Workshop on Deep Learning (2014)
6. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In: Proceedings INTERSPEECH. pp. 2426–2430. ISCA, Brno, Czech Republic (August 2021)
7. Dang, T., Dimitriadis, A., Wu, J., Sethu, V., Ambikairajah, E.: Constrained dynamical neural ode for time series modelling: A case study on continuous emotion prediction. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023)
8. Duquenne, P.A., Schwenk, H., Sagot, B.: Sonar: sentence-level multimodal and language-agnostic representations. arXiv e-prints pp. arXiv–2308 (2023)

9. Dutta, S., Ganapathy, S.: Multimodal transformer with learnable frontend and self attention for emotion recognition. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6917–6921. IEEE (2022)
10. Evain, S., Nguyen, H., Le, H., Zanon Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., Besacier, L.: LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In: Proceedings INTERSPEECH. pp. 1439–1443. ISCA, Brno, Czech Republic (August 2021)
11. Ghosal, D., Majumder, N., Poria, S., Chhaya, N., Gelbukh, A.: DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 154–164. ACL, Hong Kong, China (Nov 2019)
12. Graves, A., Graves, A.: Long short-term memory. Supervised sequence labelling with recurrent neural networks pp. 37–45 (2012)
13. Hsu, J.H., Wu, C.H., Wei, Y.H.: Speech emotion recognition using decomposed speech via multi-task learning. In: Proceedings INTERSPEECH. pp. 4553–4557. ISCA, Dublin, Ireland (2023)
14. Jhin, S.Y., Kim, S., Park, N.: Addressing prediction delays in time series forecasting: A continuous gru approach with derivative regularization. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 1234–1245. ACM, New York, NY, USA (2024)
15. Joffily, M., Coricelli, G.: Emotional valence and the free-energy principle. PLoS computational biology **9**(6), e1003094 (2013)
16. Khorram, S., McInnis, M.G., Provost, E.M.: Jointly aligning and predicting continuous emotion annotations. IEEE Transactions on Affective Computing **12**(4), 1069–1083 (2021)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2017)
18. Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Schuller, B., Star, K., Hajiyev, E., Pantic, M.: SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(3), 1022–1040 (March 2021)
19. Kuppens, P.: It's about time: A special section on affect dynamics. Emotion Review **7**(4), 297–300 (2015)
20. Lim, W., Jang, D., Lee, T.: Speech emotion recognition using convolutional and recurrent neural networks. In: 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA). pp. 1–4. IEEE (2016)
21. Lin, L.I.: A concordance correlation coefficient to evaluate reproducibility. Biometrics **45**(1), 255–268 (March 1989)
22. Lu, C., Zheng, W., Lian, H., Zong, Y., Tang, C., Li, S., Zhao, Y.: Speech emotion recognition via an attentive time–frequency neural network. IEEE Transactions on Computational Social Systems **10**(6), 3159–3168 (2022)
23. Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: IEEE International conference on acoustics, speech and signal processing (ICASSP). pp. 2227–2231. IEEE (2017)
24. Parcollet, T., Nguyen, H., Evain, S., Zanon Boito, M., Pupier, A., Mdhaffar, S., Le, H., Alisamir, S., Tomashenko, N., Dinarelli, M., Zhang, S., Allauzen, A., Coavoux, M., Estève, Y., Rouvier, M., Goulian, J., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., Besacier, L.: Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech. Computer Speech Language **86**, 101622 (2024)

25. Parthasarathy, S., Busso, C.: Jointly predicting arousal, valence and dominance with multi-task learning. In: Proceedings INTERSPEECH. pp. 1103–1107. ISCA, Stokholm, Sweden (August 2017)
26. Praveen, R.G., Cardinal, P., Granger, E.: Audio–visual fusion for emotion recognition in the valence–arousal space using joint cross-attention. IEEE Transactions on Biometrics, Behavior, and Identity Science **5**(3), 360–373 (2023)
27. Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.M., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., Pantic, M.: AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition (July 2019)
28. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–8 (April 2013)
29. Sander, D.: Models of emotion. In: Armony, J.L., Vuilleumier, P. (eds.) The Cambridge Handbook of Human Affective Neuroscience, pp. 5–56. Cambridge University Press, Cambridge, UK (2013)
30. Scherer, K.R., Moors, A.: The emotion process: Event appraisal and component differentiation. Annual review of psychology **70**(1), 719–745 (2019)
31. Srinivasan, S., Huang, Z., Kirchhoff, K.: Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. In: ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6442–6446 (2022)
32. Tarantino, L., Garner, P.N., Lazaridis, A., et al.: Self-attention for speech emotion recognition. In: Interspeech. pp. 2578–2582 (2019)
33. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5200–5204 (2016)
34. Trull, T.J., Lane, S.P., Koval, P., Ebner-Priemer, U.W.: Affective dynamics in psychopathology. Emotion Review **7**(4), 355–361 (2015)
35. Tu, G., Wen, J., Liu, C., Jiang, D., Cambria, E.: Context- and sentiment-aware networks for emotion recognition in conversation. IEEE Transactions on Artificial Intelligence **3**(5), 699–708 (2022)
36. Tzirakis, P., Zhang, J., Schuller, B.W.: End-to-end speech emotion recognition using deep neural networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5089–5093 (2018)
37. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., Schuller, B.W.: Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10745–10759 (September 2023)
38. Wang, J., Mine, T.: Multi-task learning for emotion recognition in conversation with emotion shift. In: Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation. pp. 257–266. ACL, Hong Kong, China (December 2023)
39. Waugh, C.E., Kuppens, P.: Affect dynamics. Springer (2021)
40. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R.: Abandoning Emotion Classes – Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies. In: Proceedings INTERSPEECH. pp. 597–600. ISCA, Brisbane, Australia (September 2008)