Interspeech 2024
1-5 September 2024, Kos, Greece

Evaluating Transformer-Enhanced Deep Reinforcement Learning for Speech Emotion Recognition

Siddique Latif[1], *Raja Jurdak*[1], *Björn W. Schuller*[2,3]

[1]Trusted Networks Lab, Queensland University of Technology, Australia
[2]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[3]CHI – Chair of Health Informatics, MRI, Techincal University of Munich, Germany

siddique.latif@qut.edu.au

Abstract

Emotion modelling in speech using deep reinforcement learning (RL) has gained attention within the speech-emotion recognition (SER) community. However, prior studies have primarily centred around recurrent neural networks (RNNs) to capture emotional contexts, with limited exploration of the potential offered by more recent transformer architectures. This paper explores a comprehensive evaluation of training a transformer-based model using deep RL and benchmark its efficacy in SER. Specifically, we explore the effectiveness of a pre-trained Wav2vec2 (w2v2) model-based classifier within the deep RL setting. We evaluate the proposed deep RL framework using five publicly available datasets and benchmark the results with three recent SER studies using two deep RL methods. Based on the results, we show that the transformer-based RL agent not only demonstrates an improvement in SER accuracy but also shows a reduction in the time taken to begin emotion classification, outpacing the RNNs that have been commonly used to date. Moreover, by leveraging pre-trained transformers, we observe a reduced need for extensive pre-training which has been a norm in prior research.
Index Terms: speech emotion recognition, human-computer interaction, computational paralinguistics, reinforcement learning

1. Introduction

Deep Reinforcement Learning (RL) emerges as a vibrant area of research, focusing on training agents to optimize decisions and actions through interactions with an environment, aiming to maximize cumulative rewards [1]. Its popularity surges due to remarkable achievements in complex games like AlphaGo [2] and AlphaStar [3]. Beyond gaming, RL demonstrates its versatility across audio tasks, showing potential in audio enhancement [4], automatic speech recognition [5], spoken dialogue systems [6], and music generation [7]. Specifically, RL finds promising applications in speech emotion recognition (SER) [8], where, for example, an emotion detection agent in autonomous driving systems identifies potential hazards, such as road rage, by analyzing drivers' speech and expressions [9]. The agent's reward mechanism penalizes misclassifications, ensuring precise emotion recognition [10]. Our research also focuses on advancing SER through deep reinforcement learning techniques.

The exploration of deep Reinforcement Learning (RL) in SER has gained considerable attention in recent years. Several pioneering studies (e.g., [11, 12, 13]) have ventured into the realm of RL for SER, optimising Recurrent Neural Network (RNN) [14]-based agents to harness their inherent capability to learn contextual nuances from speech data. RNNs, known for their proficiency in capturing temporal dependencies within sequential data, leverage the computationally intensive back-propagation through time (BPTT) technique [15] to achieve this end. Transformers have addressed the challenges posed by the sequential processing requirements of RNNs by leveraging the self-attention mechanism, which enables the model to learn temporal relationships in input sequences effectively [16]. This architectural innovation has propelled transformers to the forefront of SER research (e.g., [17, 18, 19]), demonstrating their superior contextual representation learning capabilities and computational efficiency [18]. Previous studies predominantly employed transformers in supervised learning settings [8], where models learn from a predefined set of labelled data, without the need for interaction with the environment [20]. This trend underscores a gap in the literature, highlighting the potential for benchmarking transformers within a deep RL framework, where models could dynamically adapt and optimise their performance through direct interaction with their environment.

This paper's key contribution is investigating the efficacy of transformer-based classifiers for Speech-Emotion Recognition (SER) using deep Reinforcement Learning (RL). To enhance the performance of SER, a pre-trained Wav2vec2 (w2v2) model [21] is employed as the foundation for the RL agent. We present a comprehensive analysis using two distinct deep RL methods and benchmark against three recent studies [11, 22, 12]. We replicate their configurations with our proposed RL agent across five publicly available datasets to ensure a fair comparison. The results demonstrate that pre-trained transformers lead to superior performance in (i) within-corpus emotion classification compared to [12], (ii) anger detection akin to [11], and (iii) domain adaptation as conducted in [22]. Additionally, our investigation into transformer size uncovers its significant impact on the SER performance within a deep RL context.

2. Related Work

Several studies delve into deep reinforcement learning for SER, with noteworthy approaches. Lakomkin [11] introduces EmoRL, leveraging an RNN classifier with the Monte Carlo Policy Gradient (REINFORCE) [23] to enhance classification speed and accuracy. This model, aimed at real-time robotic anger detection, identifies angry emotions from neutral ones up to 1.75 times faster, maintaining comparable recognition rates to GRU baselines that process entire utterances. Another study [13] applies deep RL to emotion classification in videos, simulating human emotional interpretation through a CNN-LSTM agent trained via RL, demonstrating improved emotion modelling. Rajapakshe et al. [12] employ an LSTM-based model with a novel Zeta policy for SER optimisation, also examining the impact of pre-trained LSTM agents on SER performance.

Previous studies [11, 22, 12] have used RNN-based agents to optimise SER using deep RL without exploring transformers-based architectures. Transformers have been a major break-

10.21437/Interspeech.2024-1827

through in the field of natural language processing (NLP) and speech processing. They use self-attention mechanisms to model long-range dependencies and achieve improved performance compared to RNNs for various speech-related tasks including speech emotion recognition [19]. Various studies use transformer-based architectures for SER. For instance, Chen et al. [24] present a key-sparse transformer for SER that aims to learn emotional information by reducing redundant information. Wagner et al. [18] conduct comprehensive evaluations on transformer-based pre-trained models for dimensional SER. Based on the results, they show that transformer-based architectures are state-of-the-art architectures for SER and they are more robust to small noise perturbations in contrast to a CNN-based baseline. None of the above-mentioned studies have explored transformers in RL settings in SER. Therefore, this paper explores the use of a transformer-based emotion classifier optimised using deep RL. Various other studies [25, 26, 19] also explored transformer-based architectures for SER, however, none of the studies explores the optimisation of transformers based agent for SER using the deep RL.

3. Methodology

This section presents the functioning of our deep reinforcement learning (RL) framework, featuring a transformer-based classifier within the agent module.

3.1. Transformer-based Agent

The transformer network consists of a w2v2 classifier as shown in figure 1. The network learns from raw speech and generates an output for a particular input in an RL setting. We select a transformer-based framework for this task inspired by [27, 18]. Our proposed framework builds on top of Wav2vec 2.0 (w2v2) [21] and we use a simple classification head for SER. The proposed model is shown in Figure 1, where we apply average pooling over the hidden states of the last transformer layer and feed the result through a hidden layer and a final output layer (the pooled embeddings and the hidden layer outputs are dropped out). During the deep RL optimisation, we only fine-tuned the transformer layer with the Adam optimiser. We use two RL algorithms including REINFORCE and Q-learning for comparison with previous studies.

3.2. Deep Reinforcement Learning

In reinforcement learning (RL), the 'agent' and 'environment' constitute core elements. The agent's decisions result in 'actions,' triggering environmental responses in the form of rewards and new states. We employ a transformer-based classifier within the agent module for informed decisions. To formalise this scenario, we adopt the Markov decision process (MDP). In the SER problem, predicted classes correspond to actions (a), and states (s) encompass batched audio samples (η). The pivotal role of the RL agent is to make decisions that lead to actions (a)), which consequently yield rewards (r_t) based on a reward function. This function assigns positive values for actions that align with the ground truth class (g_t) and negative values, otherwise. The agent's action probabilities stem from a transformer-based network, adept at capturing data relationships. The core goal remains to refine decision strategies for maximum returns. This is achieved through a policy $\pi(a_t, s_t)$ guiding action selection given observed state s_t.

In the context of approximating gradients to optimise our objective function, we evaluate the REINFORCE algorithm and

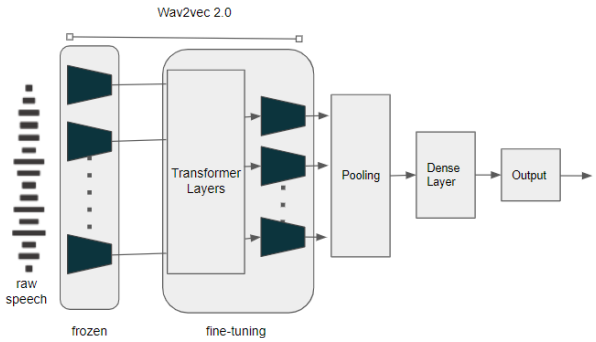

Figure 1: *Transformer-based model architecture. During the RL optimisation, we only update the transformer layers. This architecture is used for both within-corpus and cross-corpus evaluations using deep RL algorithms.*

Q-learning. The REINFORCE algorithm operates by iteratively adjusting the policy based on the accumulated rewards received during the agent's interactions with the environment. The process starts with the environment sending the initial state, s_1, to the RL agent. The agent employs the policy network, which in our case is a transformer-based classifier, to compute action probabilities and selects the action a_1 with the highest probability. Subsequently, the environment calculates the reward r_1 for the action-state pair and sends it, along with the next state s_{t+1}, back to the agent. These exchanges of values (r_i, a_i, and s_i) are stored to train the policy network within an episode. Implementing the REINFORCE algorithm with our proposed transformer-based policy network allows for a thorough comparison and benchmarking against the RNN-based model used in [11], thereby enhancing the context of our study.

In contrast, Q-learning equips the RL agent with a policy $\pi(a, s)$ indicating the probability of taking action a given state s. The Q-value, also referred to as the quality value, predicts the expected reward upon executing action a within state s. Notably, Q-learning employs a deep neural network named a Deep Q Network (DQN) to approximate Q-values, which becomes especially significant in deep Q-learning. For our research context, we adopt a transformer-based model as the DQN to approximate Q-values. This choice enables us to apply Q-learning for cross-corpus evaluation (Section 5.2), ensuring fair comparisons and benchmarking against [22].

4. Experimental Setting

4.1. Datasets Details

We selected four publicly available popular emotional datasets for SER evaluations. The details of these datasets are presented below.

IEMOCAP: The IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset [28] is a multimodal corpus featuring English dyadic conversations. The utterances in IEMOCAP are annotated by 3-4 evaluators across 10 emotional categories. In line with prior research [11, 12, 22], our experimentation utilises the same number of samples for four emotions.

SAVEE: The SAVEE database was captured from four male native English speakers, aged between 27 and 31 years. This collection encompasses utterances annotated with 8 emotions. Yet, to maintain consistency with [12] and fair comparison, in this work, we utilise four emotions: happiness, sadness, anger, and neutral.

ESD: In our cross-corpus evaluations, we select the Emo-

tional Speech Dataset (ESD) [29]. ESD is an open-source emotional speech data that incorporates utterances from 20 speakers across two languages (Mandarin and English), categorised into five emotion classes: happy, surprised, neutral, angry, and sad. In this work, we perform a comparison with [22] for cross-corpus SER using the English part for four emotions including happy, sadness, anger, and neutral.

EmoDB: EmoDB [30] is an emotional corpus in the German language. 10 actors (comprising 5 males and 5 females) recorded 10 scripted texts across 7 distinct emotions. For our cross-language investigations, we use EmoDB as the target dataset in comparison with [22], therefore, we use four emotion classes: anger, happiness, sadness, and neutral.

MSP-IMPROV: This corpus is also a multimodal emotional database recorded from 12 actors performing dyadic interactions [31]. The utterances in MSP-IMPROV are annotated in four categorical emotions: angry, happy, neutral, and sad. To be consistent with [22], we use MSP-IMPROV in cross-corpus evaluations.

4.2. Model Configuration

In our proposed RL setup, we use a w2v2-based agent for all the experiments. w2v2 contains 12 transformer blocks, hidden units dimension of 768, and 8 attention heads. This model is pre-trained in a self-supervised way on LibriSpeech [32] data that contains 960 hours of speech. One of the main goals of this work is to show the effectiveness of utilising a transformer-based model in an RL setting to improve SER performance. For all the experiments, we use raw speech as input for the w2v2-based agent.

5. Experiments and Evaluations

In this section, we present the results with our proposed setup and compare them with three different studies [12, 11, 22]. Results are presented below for within-corpus and cross-corpus.

5.1. Within Corpus Benchmark

In this experiment, we present SER results in within-corpus settings and compare the results with [12, 11]. In [11], the authors used the gated recurrent units (GRUs) based model in an RL setting to determine the earliest reasonable time to classify an emotion from the given input speech. They use the IEMOCAP corpus and perform binary classification for anger and neutral speech detection. They consider REINFORCE either to terminate or wait for the next frame of the speech utterance. The termination action triggers the emotion classifier's decision which can be either neutral or angry. On the other hand, the wait action does not trigger the decision but waits for the next frame.

We implement the same setup with similar evaluation configurations for computing the results with the proposed w2v2-based agent. Results are presented in Table 1, which shows that the proposed model shows improved results both in performance and relative latency. Table 1 shows that EmoRL achieves 84.8 % of accuracy by reading only 0.55 % of average given speech utterance compared to the GRU baseline [33] that achieves slightly better results by processing the full given utterance. In contrast, our proposed transformer-based RL setup achieves improved performance while reducing the latency of anger detection from the given speech input. We further extend our comparison with within-corpus SER by including another study in our experiments [12]. In [12], the authors use a CNN-LSTM as an RL agent and optimise the model for SER. They pre-trained the agent (CNN-LSTM) to minimise the warm-up period and show that pre-training helps reach maximum performance and reduces the training time on the IEMOCAP and SAVEE datasets. In contrast to this study, we use a pre-trained w2v2-based model and compare the results. In order to have a fair comparison, we follow [12] by the same evaluation strategies and compute the results for both the IEMOCAP and SAVEE datasets in four emotion class SER. Table 2 shows the comparison of the results on two public datasets. It can be noted that the proposed framework based on pre-trained w2v2 achieved improved results compared to the CNN-LSTM model. In [12], the authors also perform pre-training of the CNN-LSTM model and present the SER results with 700 000 iterations. In contrast, we achieved improved results with 200 000 iterations. This shows that utilising pre-trained transformers has an edge over performing pre-training of CNN-LSTM during the deep RL training.

Table 1: *Results comparison with EmoRL [11] on IEMOCAP dataset for anger detection.*

Model	Accuracy (%)	Relative Latency
GRU Baseline [33]	85.1±3.9	1
EmoRL [11]	84.8±4.3	0.55±0.2
This work	**86.1±3.2**	**0.48±0.2**

Table 2: *Comparing within-corpus results using the proposed model with CNN-LSTM.*

Paper	Datasets	
	IEMOCAP	SAVEE
Rajapakshe et al. [12]	54.29 ± 2.50	68.90 ±0.61
This work	**56.12 ± 1.80**	**70.30 ±0.72**

5.2. Cross-corpus Benchmark

We utilise the IEMOCAP, MSP-IMPROV, and ESD corpora in cross-corpus SER and perform different experiments. We consider [22] for comparison, as no other study found in the literature for cross-corpus SER is realising deep RL. We closely follow [22] for evaluation and results computations. In our first experiment, we use the IEMOCAP and MPS-IMPROV corpora as source datasets and ESD is chosen as target data.

In [22], the authors pre-trained the CNN-LSTM-based network using the source data and the pre-trained model's parameters are transferred onto the deep RL agent. They apply a deep Q-learning algorithm to optimise the model against the target ESD dataset. For cross-corpus evaluations, RL serves as an emotion-recognising game where an RL agent takes action a and recognises the correct emotion for a state s (given an audio utterance). A reward r is calculated by the environment. For this setting, the reward is calculated by comparing the inferred emotion with ground truth by the RL agent during training with labelled datasets. A policy π is learnt by the RL agent to maximise the reward gained at each episode. In [22], the authors pre-train the model with source data and then optimise it for the target dataset using RL. We follow the same setting in evaluation strategies except that our deep RL agent is a transformer-based classifier. Results show that the proposed agent achieves improved results for both cross-corpus and cross-language settings compared to the CNN-LSTM-based RL framework used in [22].

We also extended our experimentation by showing the ef-

Table 3: *Cross-corpus results comparison with [22]. Models are trained in deep RL settings without using labels.*

Source data	Target data	UAR (%)	
		Rajapakshe et al. [22]	Proposed
IEMOCAP	MSD	63.99 ± 1.95	**65.16 ± 1.34**
MSP-IMPROV	MSD	66.10 ± 0.84	**68.43 ± 0.90**
IEMOCAP	EMODB	73.17± 0.62	**75.02± 0.58**
MSP-IMPROV	EMODB	65.50± 0.71	**67.78± 0.69**

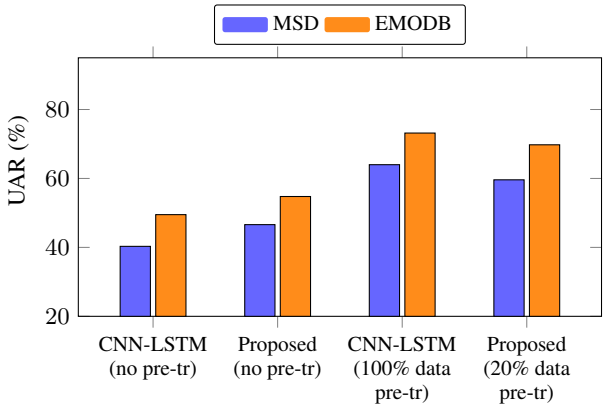

Figure 2: *Effect of pre-training (pre-tr) on the performance of deep RL for SER. Results are compared with the CNN-LSTM optimized with RL as used in [22].*

fect of pre-training on the SER performance by comparing the results with [22]. We implement a CNN-LSTM-based agent by following [22]. Overall, the CNN-LSTM approach consists of 2D convolution layers of filter sizes 5 and 3 with batch normalisation, one LSTM layer of 15 LSTM units, and one fully connected layer of 256 units. We train both the CNN-LSTM and the proposed w2v2-based agents without pre-training the source data. Results are presented in Figure 2, which shows that the proposed agent can achieve improved results compared to the CNN-LSTM approach without pre-training. We also compute the results with 20 % of source data pre-training of the proposed agent (see Figure 2). Results are comparable to the CNN-LSTM pre-trained on 100 % of source data. This shows that utilising the pre-trained transformer-based architecture as an agent helps minimise the source pre-training in the deep RL setting.

5.3. Size of Transformer

In this experiment, we explore the effect of transformer size on the performance of SER. In particular, we finetune two pre-trained models including w2v2-base and w2v2-large using REINFORCE algorithm across both within-corpus and cross-corpus settings. The performance evaluation focused on the IEMOCAP and MSD datasets, aiming to ascertain the impact of model architecture size on SER effectiveness. The comparative analysis of w2v2-b and w2v2-L models demonstrates that performance in SER tasks does not linearly correlate with the size of the architecture (see Figure 3). Specifically, both models showcase similar performance (UAR %) both in within-corpus and cross-corpus settings. These results suggest that larger model architectures do not automatically guarantee improved SER task performance. The efficiency and effectiveness of DRL in SER seem more influenced by model tuning, pre-training data specificity, and adaptability to diverse emotional contexts than just model size.

Motivated by previous studies [34, 18] in supervised train-

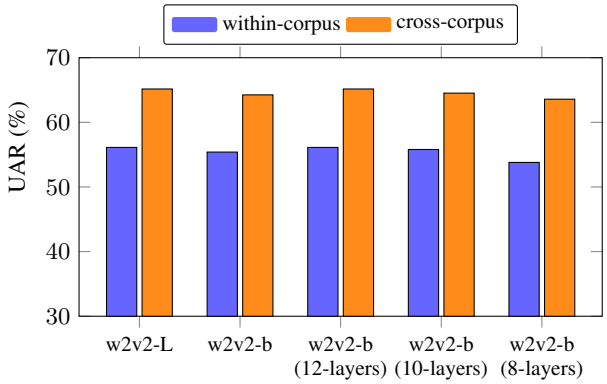

Figure 3: *Impact of transformer size on SER performance (UAR %) using DRL across within-corpus (IEMOCAP) and cross-corpus (IEMOCAP and MSD).*

ing, we extended our investigation using deep RL to explore the effects of reducing the transformer layers within the w2v2-base model architecture. This analysis aims to determine when layer reduction starts to negatively impact SER performance, focusing on the balance between model complexity and effectiveness. Through empirical analysis, it is observed that the w2v2-b model's performance remains stable even with fewer layers, effectively preserving its effectiveness up to a specific limit. However, a noticeable decline in performance is only observed when the model's complexity is reduced to fewer than 8 layers, as highlighted in Figure 3. The findings reveal that the transformer's size and layer count significantly affect SER performance, but not straightforwardly. The model maintains effectiveness despite reduced complexity up to a limit, underscoring the role of pre-training data quality and adaptability to diverse emotions. This study suggests optimising model development strategically, focusing on these aspects beyond mere architectural size.

6. Conclusions and Future Work

This study has successfully demonstrated the efficacy of integrating transformer-based reinforcement learning (RL) approaches within the speech emotion recognition (SER) domain. By leveraging the advanced capabilities of a pre-trained w2v2 model, we have showcased significant improvements in SER performance over traditional methods such as GRUs and CNN-LSTM networks within RL frameworks. Our findings underline the transformative potential of employing transformer-based architectures combined with RL techniques to refine and optimise emotional recognition in speech. In addition, our investigation into the impact of transformer size on SER performance revealed that strategic adjustments to the model's architecture significantly influence its efficiency and accuracy, offering key insights into the optimal design of transformer-based classifiers for SER tasks. These findings not only demonstrate the potential of combining transformer architectures with RL techniques but also emphasise the importance of model size optimisation in enhancing SER applications. Moving forward, we aim to delve into the application of RL from human feedback within this framework, seeking to further refine and elevate the capabilities of SER systems.

7. References

[1] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, M. Van Den Driessche *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[3] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[4] Y.-L. Shen, C.-Y. Huang, S.-S. Wang, Y. Tsao, H.-M. Wang, and T.-S. Chi, "Reinforcement learning based speech enhancement for robust speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6750–6754.

[5] C. Chen, Y. Hu, Q. Zhang, H. Zou, B. Zhu, and E. S. Chng, "Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 607–12 615.

[6] G. Weisz, P. Budzianowski, P.-H. Su, and M. Gašić, "Sample efficient deep reinforcement learning for dialogue systems with large action spaces," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2083–2097, 2018.

[7] N. Jiang, S. Jin, Z. Duan, and C. Zhang, "Rl-duet: Online music accompaniment generation using deep reinforcement learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 710–718.

[8] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, "A survey on deep reinforcement learning for audio-based applications," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2193–2240, 2023.

[9] M. Braun, F. Weber, and F. Alt, "Affective automotive user interfaces–reviewing the state of driver affect research and emotion regulation in the car," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–26, 2021.

[10] X. Huang, M. Ren, Q. Han, X. Shi, J. Nie, W. Nie, and A.-A. Liu, "Emotion detection for conversations based on reinforcement learning framework," *IEEE MultiMedia*, vol. 28, no. 2, pp. 76–85, 2021.

[11] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "Emorl: continuous acoustic emotion classification using deep reinforcement learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4445–4450.

[12] T. Rajapakshe, R. Rana, S. Khalifa, J. Liu, and B. Schuller, "A novel policy for pre-trained deep reinforcement learning for speech emotion recognition," in *Proceedings of the 2022 Australasian Computer Science Week*, 2022, pp. 96–105.

[13] T. Yuan and Y. Yuan, "Video emotional classification based on deep reinforcement learning," in *2023 3rd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, 2023, pp. 168–171.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[15] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[16] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: a survey," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 6778–6786.

[17] Y. Gao, L. Wang, J. Liu, J. Dang, and S. Okada, "Adversarial domain generalized transformer for cross-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2023.

[18] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[19] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, and J. Qadir, "Transformers in speech processing: A survey," *arXiv preprint arXiv:2303.11607*, 2023.

[20] P. Ladosz, L. Weng, M. Kim, and H. Oh, "Exploration in deep reinforcement learning: A survey," *Information Fusion*, vol. 85, pp. 1–22, 2022.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[22] T. Rajapakshe, R. Rana, and S. Khalifa, "Domain adapting speech emotion recognition modals to real-world scenario with deep reinforcement learning," *arXiv preprint arXiv:2207.12248*, 2022.

[23] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.

[24] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[25] I. Zenkov, "Transformer-cnn emotion recognition," https://github.com/IliaZenkov/transformer-cnn-emotion-recognition, 2021.

[26] Y. Li, Z. Li, Z. Zhang, X. Li, and J. Li, "Speech emotion recognition transformer: A novel end-to-end model for ser," *Neurocomputing*, vol. 454, pp. 1–10, 2021.

[27] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[28] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[29] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[30] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[31] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[33] C.-W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition." in *Interspeech*, 2016, pp. 1387–1391.

[34] H. Sajjad, F. Dalvi, N. Durrani, and P. Nakov, "On the effect of dropping layers of pre-trained transformer models," *Computer Speech & Language*, vol. 77, p. 101429, 2023.