Domain Adapting Speech Emotion Recognition models to real-world scenario with Deep Reinforcement Learning

Thejan Rajapakshe, Rajib Rana, Member, IEEE and Sara Khalifa Member, IEEE

Abstract—Deep reinforcement learning has been a popular training paradigm as deep learning has gained popularity in the field of machine learning. Domain adaptation allows us to transfer knowledge learnt by a model across domains after a phase of training. The inability to adapt an existing model to a real-world domain is one of the shortcomings of current domain adaptation algorithms. We present a deep reinforcement learning-based strategy for adapting a pre-trained model to a newer domain while interacting with the environment and collecting continual feedback. This method was used on the Speech Emotion Recognition task, which included both cross-corpus and cross-language domain adaption schema. Furthermore, it demonstrates that in a real-world environment, our approach outperforms the supervised learning strategy by 42% and 20% in cross-corpus and cross-language schema, respectively.

Index Terms—Reinforcement Learning, Speech Emotion Recognition, Domain Adaptation

1 INTRODUCTION

Motion-aware interaction can significantly improve human–computer interactions. Consequently, there has been much research on automatic speech emotion recognition (SER). While SER within the same corpus is possible with reasonable accuracy, cross-corpus SER performance needs significant improvement [1], [2], [3]. Cross-corpus SER is crucial to achieving emotion-aware interactions in real-life applications, as speech signals obtained from different devices, recording background, spoken language, and acoustic signal conditions can be different with the training dataset and real-world implementation [2].

Researchers have explored domain adaptation techniques to improve the cross-corpus SER performance [4], [5], [6], [7]. Domain adaptation is a transfer learning approach where a trained model undergoes training aiming to optimise for a different distribution other than the previously trained distribution. The key limitation of the existing domain adaptation methods is that they cannot be adapted to a real-life setting. In such a setting, it will be beneficial if an intelligent agent interacting with a customer/user could dynamically update itself when it wrongly categorises speech emotion. In this paper, we address this challenge. We develop a domain adaptation technique based on reinforcement learning, where we exploit the dynamically updating feature of RL to develop a cross-corpus SER technique that can be dynamically updated. A possible use case of the proposed technique is shown in figure 1. The model gets feedback from the customer regarding its interpretation of emotion in the customer’s voice. This can, for example, be implemented by a thumbs up or down button on the customer console.

Using widely used speech corpora, we evaluate the model in cross-corpus and cross-language scenarios and show that we achieve better SER performance than fully supervised benchmark models. Moreover, we create a real-life scenario and show that our proposed RL model significantly outperforms fully supervised benchmark models.

Fig. 1. Overall Architecture of using reinforcement learning in Domain Adaptation for speech emotion recognition task

2 LITERATURE REVIEW

The advancement of deep learning has been held captive due to insufficient computing capability. Since deep learning is a data-centric machine learning approach, with the evolution of the computational capability of modern-day computers, machine learning researchers were able to
train deep neural networks (DNNs) for numerous kinds of applications [8], [9]. Many Machine Learning (ML) researchers tend to work on deep learning creating a trend in deep learning and Convolutional Neural Networks (CNNs) after the ImageNet competition in 2012 and CNNs made much success in research in computer vision [10], [11], [12]. Similarly, deep learning has outperformed conventional ML algorithms in many disciplines such as video understanding [13], [14], [15], optical character recognition [16], [17], machine translation [18], [19], image generation [20], [21], [22], game playing [23], [24], [25], image enhancement [26], [27], speech recognition [28], [29], [30], image to image translation [31], [32], speech synthesis [33], [34], [35], [36], [37], etc.

Deep Reinforcement Learning, a hybrid of reinforcement learning and deep neural networks, has grown in prominence as deep learning has advanced. Deep RL has mostly been applied in applications like as gaming [25], [38], [39], recommendation systems [40], and robotics [41]. With the popularity of deep RL, the use of RL for speech-based applications has increased [42].

2.1 Domain Adaptation

Domain adaptation was introduced to overcome the ‘corpus bias’ issue. Since deep learning methods have a good reputation for extracting non-linear features in the input, domain adaptation can easily be implemented on deep learning based platforms which yields more robust and performing models [43], [44].

Unsupervised domain adaptation uses an unlabeled target dataset and applies a transformation between the domains. The latent space difference between the distributions of source and target data is minimised [53], [54] and generates a path in which the domain is transformed between source and target domains [53], [55].

Mao et al. experimented on a two-class task for domain adaptation by sharing priors between source and target domains [4]. They pre-trained a two-layer neural network with unsupervised learning and then share the common classifier parameters between source and target. Our approach follows the reinforcement learning paradigm for domain optimisation instead of unsupervised learning and we pre-train the source model with a labelled source dataset. Gharib et al. proposes unsupervised adversarial domain adaptation and pre-train the model with two conditional sets and achieved nearly 10% increase in accuracy [56].

Ahn. Y et al. proposed Few-shot learning based methodology in unsupervised domain adaptation for cross-corpus speech emotion recognition tasks [52]. They have used multiple corpora to optimise the emotion recognition robustness to unseen samples.

2.2 Reinforcement Learning for Domain Adaptation

The Reinforcement learning agent gains their knowledge via experiences and these experiences are obtained by exploring and exploitation. RL agent uses the learnt policy to choose the best action which is called exploitation. Since the RL agent does not have enough experience to update the policy in the initial phase, it explores the environment. Exploration is done by randomly selecting actions without considering policy and the experience is stored in the memory buffer with a reward. The main agitation in this exploration is the unsafe choice of actions in some states [51].

Prior knowledge can be fed to the RL agent before releasing to the real world in many ways. Demonstration based learning [57], [58], pre-training and domain adaptation [51] are some such methods to nourish the RL agent with knowledge. This reduces the amount of exploration time and the risk of unsafe exploration. Previous studies also have shown that pre-training has reduced the training time to achieve higher performance [59], [60], [61].

Koo et al. consider the feature extractor for the target domain as a generator where the parameters are similar to the latent feature set in the source and are trained with adversarial training loss. They proposed this to enhance the coherent training of the target domain feature extractor [50]. Though their research proposes an RL based approach in Domain adaptation, they have not studied speech emotion recognition.

Leonetti et al. have introduced a new RL algorithm “DARLING” which can adapt to the new environment and improve the reliability of the decisions [45]. Hazara and Kyrki proposed a transfer approach that captures the core features of a simulated system and rationalise the dynamics then combined with incremental learning. They demonstrated their approach by applying to a robot with a basketball task and showed the target model has improved task generalisation capability than direct usage [62].

Literature consists of many studies on Domain Adaptation on different methodologies for the task of speech emotion recognition. Also, there are several studies that involved reinforcement learning for speech emotion recognition tasks such as EmoRL [47]. But there is very limited research in the literature that uses Reinforcement learning in domain adaptation for speech emotion recognition. We, in this study, are observing the possibility of using RL in domain adaptation in the arena of speech emotion recognition. A concise summary of the literature on RL for domain adaptation and speech emotion classification is shown in Table 1.

3 METHODOLOGY

3.1 Reinforcement Learning in the domain of Speech Emotion Recognition

Reinforcement Learning follows the principle of behaviourist psychology and learns similar to a child learns to perform a new task. There are 2 main components in a RL problem namely, Environment and Agent. The agent performs actions on an environment and the environment returns a reward regarding the action performed along with the next state of the environment. Figure 2 shows the concise architecture of RL methodology.

An RL problem can be interpreted by the famous Markov Decision Process which has the parameters state space $S$, action space $A$ and reward $R$. A state $s \in S$ is the representation of the current environment, action $a \in A$ is a single task that can be performed on the environment, reward $r \in R$ is feedback value returned from the environment after executing the task $a$ for the state $s$. RL agent learns a policy $\pi(a,s)$ that represents a probability of taking action $a$ for a specific state $s$. Q-value or quality value
TABLE 1
Summary and focus on the literature on RL for domain adaptation and speech emotion classification

| Paper                        | Reinforcement Learning | Domain Adaptation | Speech Emotion Recognition |
|------------------------------|------------------------|-------------------|---------------------------|
| Leonetti et al, 2016 [45]   | ✓                      | ✓                 | ✓                         |
| Mao et al, 2016 [4]         | ✓                      | ✓                 | ✓                         |
| Carr et al, 2018 [38]       | ✓                      | ✓                 | ✓                         |
| Lakomkin et al, 2018 [47]   | ✓                      | ✓                 | ✓                         |
| Hossain and Muhammad, 2019 [49] | ✓              |                   | ✓                         |
| Koo et al, 2019 [50]        | ✓                      | ✓                 | ✓                         |
| Arndt et al, 2020 [51]      | ✓                      | ✓                 | ✓                         |
| Ahn et al, 2021 [52]        | ✓                      | ✓                 | ✓                         |
| This paper                  | ✓                      | ✓                 | ✓                         |

Fig. 2. Deep Q network in Reinforcement Learning architecture

represents the expected reward when action $a$ is performed on an environment with state $s$. Optimum Q-value $Q^*(a, s)$ can be written as in Equation 1:

$$Q^*_\pi(s, a) = \max_{\pi} E[ R_{t+1} + \gamma Q_{\pi}(S_{t+1}, A_{t+1}) ]$$ (1)

Q-values are estimated using Q-tables in q-learning while a deep neural network called Deep Q Network (DQN) is used to estimate the Q values in deep q-learning. Loss function in Equation 2 is used to calculate the loss and $Q_{\text{target}}$ can be as of Equation 3.

$$L = \sum (Q_{\text{target}} - Q)^2$$ (2)

$$Q_{\text{target}} = R(s_{t+1}, a_{t+1}) + \gamma Q(s_{t+1}, a_{t+1})$$ (3)

Combining the two equations 2 and 3, loss function can be re-written as in Equation 4.

$$L = \sum (R(s_{t+1}, a_{t+1}) + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))^2$$ (4)

Loss $L$ should be minimised in the training phase of RL.

In this study, RL is practised as an emotion recognising game where an RL agent attempts to recognise the correct emotion (action $a$) for a given audio utterance (state $s$). A reward $r$ is returned by the environment. The reward value is obtained either by the feedback from the user in a real-world scenario or by comparing ground truth and inferred emotion by the RL agent when training with labelled datasets. A policy $\pi$ is learnt by the RL agent to maximise the reward gained at each episode. The state space $S$ is defined as the distribution of speech audio, Action space $A$ is defined as the discrete emotion class and action selection is the inference of emotion to a given audio utterance. According to Equation 4 to minimise $L$, $Q(s_t, a_t)$ should be maximised whereas Reward $R_{t+1}$ should be maximum at $L^*$.

Pre-training the DQN model prior to utilising an RL agent has proven to increase the performance of the model [60], [61]. Here, RL is used to optimise the model which is pre-trained by the source domain to the target domain. RL agent with a pre-trained DQN model interacts with the environment and tries to optimise the DQN to the target domain.

This study is to observe the possibility of using reinforcement learning to optimise a pre-trained model to a target domain without using labelled datasets as target data.

Though Reinforcement Learning methodologies do not report performance based on accuracy since there is no labelled data, we are using a supervised learning approach as the baseline to compare the performance in our methodology. We use 4 labelled datasets and section 4.1 describes about the datasets used in this study. A subsection of a size 20% of each dataset is reserved as a testing dataset and the remaining subsection is used for either pre-training or RL.
domain optimisation. Section 4 describes about the different kinds of experiments carried out within this study to see the effectiveness of our approach.

3.2 Feature Extraction

We use Mel Frequency Cepstral Coefficients (MFCC) as the input features to models in this research since MFCCs are popular features used in speech audio analysis [63], [64]. We use 2,048 as the frame length and hop length as 512 and extracted 40 MFCCs using Librosa python library for audio and music analysis [65].

3.3 Model Architecture

The functionality of the DNN model is to classify emotion embedded in speech utterances and the study uses MFCCs of speech utterances as the input for the model. We use CNN and Long Short-Term Memory (LSTM) combined model since it motivates the capability to learn both frequency and temporal components in the speech signal [61], [66]. We stack CNN, LSTM and fully connected layers respectively for the model to learn the discriminative features [64].

As reflected in figure 2, the feature matrix is fed into two layers of 2D convolution layers of filter sizes 5 and 3 respectively and the output of the first 2D convolution layer is Batch normalised. Output from the second 2D convolution layer is passed to the LSTM layer of 16 cells after flattening and then to a fully connected layer of 256 units. A dropout of 0.3 rate is applied before the last Dense output layer. The number of units in the output layer is equal to the number of emotions to be classified, in this study, four. Linear activation function of the output layer is used when the model is used in the RL agent since it outputs Q-values for a specific state and they should not be normalised. The input shape of the model is 40 × 87, where 40 MFCCs are used in the input and 87 is the number of MFCC frames.

The model is optimised using Adam Optimiser with a learning rate of $2.5 \times 10^{-4}$. We use the popular Deep Learning API Keras [67] with Tensorflow [68] as the backend for modelling and training purposes in this study.

4 Experimental Setup

4.1 Datasets

To measure the performance of the proposed RL based Domain Adaptation (RL-DA) approach, we use 4 popular datasets for our experiments namely; MSP-IMPROV [69], IEMOCAP [70], ESD [71] and EmoDB [72]. The selection of datasets covers the ability to use them for Cross-Corpus (CC) experiments as well as Cross-Language (CL) experiments in the field of speech emotion recognition.

Since the datasets do not contain a balanced number of utterances under each emotion class, we use an audio augmentation technique to generate missing audio utterances. We utilise Vocal Tract Length Perturbation (VLTP) [73] to augment the missing audio utterances.

4.1.1 IEMOCAP (IEM)

A popular dataset containing 12 hours of multi-speaker acted audio-visual data. IEMOCAP dataset contains dyadic sessions with both improvised and scripted scenarios. It contains both categorical such as happiness, sadness, anger and neutrality as well as dimensional labels such as valence, dominance and activation [70]. This study uses the audio data modality from improvised scenarios and categorical labels.

4.1.2 MSP-IMPROV (MSP)

An acted audio-visual database that is popular in multimodal speech emotion recognition research. The database was initially developed for a study in audio-visual emotional perception, but this dataset has also been used in speech emotion recognition studies [52], [74], [75], [76] which makes it a candidate for our study as well. Dataset was recorded in a controlled environment and consists of 20 determined scripts covering the major 4 emotions; happiness, sadness, anger and neutral [69].

4.1.3 Emotional Speech Dataset (ESD)

A public database of speech data that was initially developed for speech synthesis and voice conversion. It contains utterances from 20 speakers in two languages (10 Mandarin and 10 English) categorically labelled into 5 emotions such as happy, surprise, neutral, angry and sad [71]. We use utterances from 10 English speakers under 4 emotions happiness, sadness, anger and neutral.

4.1.4 Berlin EmoDB (EmoDB)

An actor recorded database of emotional speech utterances in the German language. 10 actors (5 male and 5 female) in age range between 21 years and 35 years were used to speak 10 scripted texts in 7 different emotions such as anger, boredom, disgust, fear, happiness, sadness and neutral [72]. This study uses EmoDB as the target dataset for cross-language experiments containing only the utterances of emotions anger, happiness, sadness and neutral.

4.1.5 Diverse Environments Multichannel Acoustic Noise Database (DEMAND)

DEMAND dataset is a popular real-world background noise dataset. The database contains background noises from 18 different environments under 6 categories. An array of 16 microphones is used to record the audio in each environment and stored in 16 channels. DEMAND dataset contains versions of both 16kHz and 48kHz sampling rates and we use 48kHz and down-sampled to 22kHz to match with other datasets used in this study.

4.2 Experiments

We run experiments to observe the behaviour of the proposed Reinforcement Learning based Domain Adaptation approach under scenarios: 1. Pre-train with source dataset and domain optimising with target dataset separately. 2. Pre-train with the subset of source dataset and domain optimising with a mix of the remaining source and target datasets. 3. Following a real-world scenario on domain adaption.
4.2.1 Experiment with separate datasets

Here, we examine the behaviour of the proposed Reinforcement Learning based Domain Adaptation approach for separate source and target datasets. We select datasets considering two schemas, cross-corpus and cross-language. We use IEMOCAP and MSP-Improv datasets as source datasets while the ESD dataset is used as the target dataset in the cross-corpus schema and EmoDB is used as the target dataset in cross-language schema. Table 2 show how the datasets are used in each experiment.

| Training Dataset Source | Target | Testing Dataset |
|-------------------------|--------|-----------------|
| CC                      | IEM    | IEM + ESD       |
|                         | MSP    | MSP + ESD       |
| CL                      | IEM    | IEM + EmoDB     |
|                         | MSP    | MSP + EmoDB     |

First, the modal is pre-trained using the training subset of the source dataset and the trained modal parameters are transferred onto the DQN of the RL agent. Then the RL deep Q learning algorithm is executed to optimise the DQN modal in the environment. RL optimised DQN modal parameters are then applied to a DNN in a supervised learning set-up and tested with a training subset of both source and target datasets. We used Max Boltzman Policy (Max.B) [77] as the RL policy in RL agent.

Supervised learning approach based performance is also measured to compare the performance of our methodology as the baseline (Baseline 1). We create a DNN modal that is similar in architecture as used in DQN and is first pretrained by the source dataset. Then the same modal is trained by the target dataset and measured the accuracy by testing with the testing subset of both source and target datasets.

4.2.2 Experiment by mixing source and target datasets

This experiment carried on to see the behaviour of our proposed methodology if we include a subset of the source dataset in the RL optimisation. The ambition of such mixing is to preserve some features of the source dataset than totally deviating from the source dataset.

This experiment is performed similar to the previous experiment mentioned in 4.2.1, but with the following amendments. A subset of the training subset of source dataset is used in pre-training and a mix of the remaining training subset of the source dataset and training subset of the target dataset is used in RL optimisation. Table 3 shows the dataset used in this experiment.

| Training Dataset Source | Target | Testing Dataset |
|-------------------------|--------|-----------------|
| CC                      | 50% IEM | IEM + ESD       |
|                         | 50% MSP | MSP + ESD       |
| CL                      | 50% IEM | IEM + EmoDB     |
|                         | 50% MSP | MSP + EmoDB     |

We measure the performance by using supervised learning as a baseline (Baseline 2) to compare with the proposed approach. First, the modal is pre-trained with a subset of the training subset of the source dataset and then trained with a mix of the remaining training subset of the source dataset and target dataset as mentioned in Table 3.

4.2.3 Experiment by following a real-world scenario

This experiment is conducted to observe the performance difference between supervised learning and RL optimisation in a real-world scenario. Since real-world scenarios do not have the possibility to label the dataset, baseline performance is measured after only pre-training with the source dataset.

The methodology of RL optimisation is similar to the experiment described in 4.2.1. The baseline modal is first trained with a training subset of the source dataset and then testing accuracy is measured with testing subsets of both source and target datasets (Baseline 3).

We also measured the performance of our strategy with real-world audio. We mixed background noise audio with our target datasets to create a new variation of the dataset with background noise. Audio from the kitchen environment of the DEMAND dataset is used as the background noise to mix with target datasets (ESD and EmoDB).

5 Evaluations

This section reports the results and description obtained through the experiments in section 4.

5.1 Experiment with separate datasets

We calculate the mean testing accuracy of the model after training with separate datasets as mentioned in section 4.2.1. The accuracy of the baseline (Baseline 1) model after training with supervised learning and the accuracy of model after training with RL-DA approach (RL-DA) are reported in the table 4.

| Dataset | Source | Target | Accuracy |
|---------|--------|--------|----------|
| CC      | IEM    | ESD    | 68.54 ± 2.26 |
|         | MSP    | ESD    | 80.16 ± 0.62 |
| CL      | IEM    | EmoDB  | 47.18 ± 4.10 |
|         | MSP    | EmoDB  | 58.40 ± 4.62 |

We find that the accuracy of RL-DA has surpassed Baseline 1 accuracy by nearly 1% in each source-target combination. Furthermore, the standard deviation of the cross-language schema has decreased from 4 to 2 comparing SL approach and RL-DA approach respectively. This outcome is noteworthy because, unlike SL approaches, which feed labelled data into the model, RL methods only provide feedback rewards that show the accuracy of inference made during the RL optimisation phase.
5.2 Experiment by mixing source and target datasets

Results of the experiment by mixing source data in to the target dataset mentioned in section 4.2.2 is tabulated in Table 5. Accuracy of RL-DA approach is compared with Supervised Learning approach as Baseline 2.

| Source | Target Dataset | Accuracy Baseline 2 | Accuracy RL-DA |
|--------|----------------|---------------------|----------------|
| CC     | IEM + ESD      | 52.19 ± 0.95        | 68.76 ± 2.48   |
|        | MSP + ESD      | 50.83 ± 2.65        | 66.88 ± 1.09   |

It is noticeable that the testing accuracy of the RL-DA approach is outperformed the Baseline 2 accuracy nearly by 10% in each experiment. Since the target dataset includes components from the source dataset, the representations of the model learned during the pre-training stage are kept while it adapts to the new domain.

We compared the current with results from section 5.1 and visualised in Figure 3

Observing the results, it is visible that mixing of source data in to target dataset yields better results in the experiments of cross-language schema, while RL based domain adaptation performs better with separate datasets in the cross-corpus schema.

To explain the above behaviour, we visualised the change for distribution when mixing datasets vs separate. We used t-SNE [78] method to visualise the representations learnt by the model at the last layer. One t-SNE visualisation of dimensional reduced feature representation for “MSP-ESD” and “MSP-EmoDB” combination in separate schema.

average log distance between the t-SNE cluster points in the cross-language separate (s) schema is clearly larger than others. Lower the distance between the points, the higher the clustering of the points which depicts that the model has learnt representations to classify the more correctly. The
in the mixed (m) schema. This shows that when trained in separate schema rather than mixed schema, cross-language models categorise emotions more sparsely.

5.3 Experiment by following real-world scenario
Testing accuracy of the models trained in three different methods (Baseline 3, RL-DA - target dataset without background noise (RL-DA w/o noise) and RL-DA - target dataset with background noise (RL-DA w/ noise)) in a real-world scenario mentioned in 4.2.3 are tabulated in Table 6. A column shows the difference in accuracy between our approach and the baseline where one has to use a modal pre-trained only in the source dataset.

Observing the A column in Table 6, it is clear that RL-based domain adaptation (RL-DA) outperforms the baseline supervised learning strategy in the real-world situation (Baseline 3). Supervised learning methodologies do not have the possibility to adapt to the deviation of domain distribution without a labelled dataset to re-train during inference time. On the other hand, RL based learning methodologies constantly receive feedback from the environment and optimise the agent model for the domain deviations on the go. Therefore we suggest that reinforcement learning is more suitable in domain adaptation to be used in real-world scenarios because of the adaptability during optimisation phase.

\[
\mu_{SNR} = \frac{\sum \mu}{\sigma}
\]

The SNR value of the MSP-ESD combination is lower, indicating that the noise ratio of utterances is large. Given lower SNR means that the signal’s noise component is greater, we may conclude that the MSP-ESD combination already has significant noise encoded in the dataset before adding background noise. Though adding noise to the training datasets enhances performance due to reduced overfitting when generalising deep learning models [79], after introducing severe noise, accuracy decreased, supporting our hypothesis about the experiment’s behaviour (2).

6 Conclusion
This study investigates the possibility of using reinforcement learning for domain adaptation in speech emotion recognition task. Evaluations if this research suggests that using RL can improve the domain adaptability on a real-world scenario comparing to a supervised learning approach. An accuracy increment of 42% in cross-corpus and 20% in cross-language schema was observable in the results when using the proposed RL-DA approach in compared to SL approach.

We propose that the main reason for improved domain adaptability in real-world setting is the ability to give feedback to RL agent during optimising period in RL paradigm. Supervised learning does not have such feedback mechanism and it requires an experts input in labelling training data.

The outcome of this study can be implemented on a real-world emotion aware application which can adapt to the users’ emotional expression under a new domain.

Fig. 6. Testing Accuracy of each experiment mimicking real-world scenario.

Figure 6 shows the comparison between mean testing accuracy of Baseline, RL-DA without noise and RL-DA with noise after executing the real-world experiments. We observe that the accuracy of experiments (1), (3) and (4) increases with addition of background noise while only in experiment (2) the accuracy reduces with background noise addition. We further study the reason for the performance drop in experiment (2) by studying the Signal-to-Noise ratio (SNR) of the datasets. We use equation 5 to calculate the mean SNR (\(\mu_{SNR}\)) of the audio utterances of each dataset with and without background noise. Where \(\mu\) is the mean of an audio signal, and the \(\sigma\) is the standard deviation of the signal. And we tabulated the mean SNR values in the Table 7.

6 References
[1] R. Milner, M. A. Jalal, R. W. Ng, and T. Hain, “A Cross-Corpus Study on Speech Emotion Recognition,” 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings, pp. 304–311, 12 2019.
[2] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, “Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies,” IEEE Transactions on Affective Computing, vol. 1, no. 2, pp. 119–131, 7 2010.
[3] J. Gideon, M. G. McInnis, and E. M. Provost, “Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDG),” IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 1055–1068, 2021.
[4] Q. Mao, W. Xue, Q. Rao, F. Zhang, and Y. Zhan, “Domain adaptation for speech emotion recognition by sharing priors between related source and target classes,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 3 2016, pp. 2608–2612.
[5] Y. Xi, P. Li, Y. Song, Y. Jiang, and L. Dai, “Speaker to emotion: Domain adaptation for speech emotion recognition with residual adapters,” 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019, pp. 513–518, 11 2019.
[6] E. N. Ocquaye, Q. Mao, Y. Xue, and H. Song, “Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network,” International Journal of Intelligent Systems, vol. 36, no. 1, pp. 53–71, 1 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/int.22291
A. Carrio, C. Sampedro, A. Rodriguez-Ramos, and P. Campoy, "A

N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu,

C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional

L. Deng and D. Yu, "Deep Learning," Foundations and Trends in

J. Lin, C. Gan, and S. Han, "TSM: Temporal Shift Module for Effi-

S. Srivastava, J. Priyadarshini, S. Gopal, S. Gupta, and H. S. Dayal, "Optical Character Recognition on Bank Cheques Using 2D Convolution Neural Network," Advances in Intelligent Systems and Computing, vol. 697, pp. 589–596, 2019.

M. Namysl and I. Konya, "Efficient, lexicon-free OCR using deep learning," Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, pp. 295–301, 9 2019.

M. T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421, 8 2015. [Online]. Available: https://arxiv.org/abs/1508.04025v5

P. Bahar, C. Brix, and H. Ney, "Towards Two-Dimensional Sequence to Sequence Model in Neural Machine Translation," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 3009–3015, 10 2018. [Online]. Available: https://arxiv.org/abs/1810.03975v1

A. Noguchi and T. Harada, "Image Generation From Small Datasets via Batch Statistics Adaptation," in IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2750–2758.

L. Ardizzone, C. Lüth, J. Kruse, C. Rother, U. Köthe, V. Learning, and L. Heidelberger, "Guided Image Generation with Conditional Invertible Neural Networks," 7 2019. [Online]. Available: https://arxiv.org/abs/1907.02392v3

M. Z. Khan, S. Jabeen, M. U. G. Khan, T. Saba, A. Rehmat, A. Rehman, and U. Tariq, "A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks," IEEE Access, vol. 9, pp. 1250–1260, 2021.

R. R. Torrado, P. Bontraguer, J. Togelius, J. Liu, and D. Perez-Liebana, "Deep Reinforcement Learning for General Video Game AI," IEEE Conference on Computational Intelligence and Games, CIG, vol. 2018–August, 6 2018. [Online]. Available: https://arxiv.org/abs/1806.02448v1

M. Kempska, M. Wydmuch, G. Runc, J. Toczek, and W. Jaskowski, "VIZDoom: A Doom-based AI Research Platform for Visual Reinforcement Learning," IEEE Conference on Computational Intelligence and Games, CIG, vol. 6, 5 2016. [Online]. Available: https://arxiv.org/abs/1507.07762v2

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," 12 2013. [Online]. Available: https://arxiv.org/abs/1312.5602v1

G. Li, Y. Yang, X. Qu, D. Cao, and K. Li, "A deep learning based image enhancement approach for autonomous driving at night," Knowledge-Based Systems, vol. 213, p. 106617, 2 2021.

Y. Guo, H. Li, and P. Zhuang, "Underwater Image Enhancement Using a Multiscale Dense Generative Adversarial Network," IEEE Journal of Oceanic Engineering, vol. 45, no. 3, pp. 862–870, 7 2020.

G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 12 2012.

L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) based speech emotion recognition," Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, AII 2013, pp. 312–317, 2013.

A. Zeyer, K. Irie, R. Schluter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018-September, pp. 7–11, 5 2018. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1616

P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 5967–5976, 11 2016. [Online]. Available: https://arxiv.org/abs/1611.07004v3

Z. Wang, X. Liu, M. Suganuma, and T. Okatani, "Cross-Region Domain Adaptation for Class-level Alignment," 9 2021. [Online]. Available: https://arxiv.org/abs/2109.06422v1

J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," arXiv, 10 2020. [Online]. Available: http://arxiv.org/abs/2010.05646

TABLE 6
Testing Accuracy of each experiment mimicking real-world scenario

| Source | Dataset | Target | Baseline 3 | RL-DA | ∆ |
|--------|---------|--------|------------|-------|---|
| CC     | IEM     | ESD    | 32.92 ± 2.84 | 65.85 ± 3.57 | 32.93 |
|        | MSP     | ESD    | 33.71 ± 1.55 | 82.04 ± 1.44 | 48.33 |
| CL     | IEM     | EmoDB  | 33.14 ± 1.96 | 52.48 ± 3.33 | 19.34 |
|        | MSP     | EmoDB  | 39.59 ± 0.54 | 59.80 ± 0.97 | 20.21 |

| Source | Dataset | Target | w/o noise | w/ noise |
|--------|---------|--------|-----------|---------|
| CC     | IEM     | ESD    | 32.92 ± 2.84 | 75.21 ± 3.08 |
|        | MSP     | ESD    | 33.71 ± 1.55 | 77.23 ± 0.88 |
| CL     | IEM     | EmoDB  | 33.14 ± 1.96 | 53.87 ± 1.54 |
|        | MSP     | EmoDB  | 39.59 ± 0.54 | 60.78 ± 2.31 |

TABLE 7
Mean SNR of the datasets with background noise and without background noise

| Source | Target | w/ noise | w/o noise |
|--------|--------|----------|-----------|
| IEM    | ESD    | 2.93 × 10⁻⁴ | 4.15 × 10⁻⁴ |
| MSP    | ESD    | 2.96 × 10⁻⁵ | 4.20 × 10⁻⁵ |
| IEM    | EmoDB  | 1.31 × 10⁻³ | 1.85 × 10⁻³ |
| MSP    | EmoDB  | 1.43 × 10⁻³ | 2.03 × 10⁻³ |
Thejan Rajapakshe received bachelor degree in Applied Sciences from Rajarata University of Sri Lanka, Mihintale and Bachelor of Information Technology from University of Colombo, School of Computing in 2016. He is currently a research scholar at the University of Southern Queensland (USQ). His research interests include reinforcement learning, speech processing and deep learning.

Rajib Rana is an experimental computer scientist. Advance Queensland Senior Research Fellow and an Associate Professor of Computer Science in the University of Southern Queensland. He is also the Director of IoT Health research program at the University of Southern Queensland. He is the recipient of the prestigious Young Tall Poppy QLD Award 2018 as one of Queensland’s most outstanding scientists for achievements in the area of scientific research and communication. Rana’s research work aims to capitalise on advancements in technology along with sophisticated information and data processing to better understand disease progression in chronic health conditions and develop predictive algorithms for chronic diseases, such as mental illness and cancer. His current research focus is on Unsupervised Representation Learning, Reinforcement Learning, Adversarial Machine Learning and Emotional Speech Generation. He received his B.Sc. degree in Computer Science and Engineering from Khulna University, Bangladesh with Prime Minister and President’s Gold medal for outstanding achievements and Ph.D. in Computer Science and Engineering from the University of New South Wales, Sydney, Australia in 2011. He received his postdoctoral training at Autonomous Systems Laboratory, CSIRO before joining the University of Southern Queensland in 2015.

Sara Khalifa is currently a senior research scientist at the Distributed Sensing Systems research group, Data61—CSIRO. She is also an honorary adjunct lecturer at University of Queensland and conjoint lecturer at University of New South Wales. Her research interests rotate around the broad aspects of mobile and ubiquitous computing, mobile sensing and Internet of Things (IoT). She obtained a PhD in Computer Science and Engineering from UNSW (Sydney, Australia). Her PhD dissertation received the 2017 John Makepeace Bennett Award which is awarded by CORE (Computing Research and Education Association of Australasia) to the best PhD dissertation of the year within Australia and New Zealand in the field of Computer Science. Her research has been recognised by multiple iAwards including 2017 NSW Mobility Innovation of the year, 2017 NSW R&D Innovation of the year, National Merit R&D Innovation of the year, and the Merit R&D award at the Asia Pacific ICT Alliance (APICTA) Awards, commonly known as the “Oscar” of the ICT industry in the Asia Pacific, among others.

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

[72] F. Burkhardt, A. Paeschke, M. Rolfs, W. Sendimeier, and B. Weiss, “A Database of German Emotional Speech,” in *Interspeech*, Lisbon, 2005, pp. 1517–1520.

[73] N. Jaitly and E. Hinton, “Vocal Tract Length Perturbation (VTLP) improves speech recognition,” in *30th International Conference on Machine Learning*, 2013. [Online]. Available: http://www.cs.toronto.edu/~hinton/absps/perturb.pdf

[74] Y. Gao, S. Okada, L. Wang, J. Liu, and J. Dang, “Domain-Invariant Feature Learning for Cross Corpus Speech Emotion Recognition,” pp. 6427–6431, 2022.

[75] X. Xie, J. Lou, and L. Zhang, “Speech Emotion Recognition Model with Time-Scale-Invariance MFCCs as Input,” *2021 International Wireless Communications and Mobile Computing, IWCMC 2021*, pp. 537–542, 2021.

[76] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks with Auditory Front-Ends,” *IEEE Access*, vol. 8, pp. 16560–16572, 2020.

[77] M. Wiering, “Explorations in Efficient Reinforcement Learning,” Ph.D. dissertation, University of Amsterdam, 1999. [Online]. Available: https://hdl.handle.net/11245/1.164863

[78] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[79] S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T. F. Zheng, and Y. Li, “Noisy training for deep neural networks in speech recognition,” *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–14, 2015.