Deep Representation Learning for Speech Emotion Recognition

A Thesis submitted by

Siddique Latif
BSEE, MSEE

For the award of

Doctor of Philosophy

2022

ABSTRACT

The success of machine learning (ML) algorithms generally depends on the quality of data representation or features. Good representations of the data make it easier to develop machine learning predictors or even deep learning (DL) classifiers. In speech emotion recognition (SER) research, the emotion classifiers heavily depend on hand-engineered acoustic features, which are typically crafted with human domain knowledge. Automatic emotional representation learning from the speech is a challenging task because speech contains different attributes of the speaker (i.e., gender, age, emotion, etc.) along with the linguistic message. Recent advancements in DL have fuelled the area of deep representation learning from speech. The prime goal of deep representation learning is to learn the complex relationships from input data, usually through the nonlinear transformations. Research on deep representation learning has significantly evolved, however, very few studies have investigated emotional representation learning from speech using advanced DL techniques. In this thesis, I explore different deep representation learning techniques for SER to improve the performance and generalisation of the systems. I broadly solve two major problems: (1) how deep representation learning can be utilised to improve the performance of SER by utilising the unlabelled, synthetic, and augmented data; (2) how deep representation learning can be applied to design generalised and robust SER systems. To address these problems, I propose different deep representation learning techniques to learn from unlabelled, synthetic, and augmented data to improve the performance and generalisation of SER systems. I found that injecting the additional unlabelled, augmented, and synthetic data in SER systems help improve the performance of SER systems. I also show that adversarial self-supervised learning can improve cross-language SER and deeper architectures learn robust generalised representation for SER in noisy conditions.

CERTIFICATION OF THESIS

I Siddique Latif declare that the PhD Thesis entitled Deep Representation Learning for Speech Emotion Recognition is not more than 100,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references, and footnotes.

This Thesis is the work of Siddique Latif except where otherwise acknowledged, with most of the contribution to the papers presented as a Thesis by Publication undertaken by the student. The work is original and has not previously been submitted for any other award, except where acknowledged.

Date: **16-08-2022**

Endorsed by:

Dr. Rajib Rana
Principal Supervisor

Prof. Ji Zhang
Associate Supervisor

Student and supervisors' signatures of endorsement are held at the University.

STATEMENT OF CONTRIBUTION

The papers produced from this doctoral Thesis are a joint contribution of student and supervisory team. However, majority of work is completed by the student. This Thesis explores deep machine learning based models to improve the performance and generalisation of speech emotion recognition systems. The details of contribution are given below:

Paper 1:

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Bjoern W. Schuller. "Survey of Deep Representation Learning for Speech Emotion Recognition," in IEEE Transactions on Affective Computing (TAC), (2021), doi: 10.1109/TAFFC.2021.3114365.

Siddique Latif contributed 75% to this paper. Collectively, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W. Schuller contributed the remainder.

Paper 2:

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Björn Wolfgang Schuller. "Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition," in IEEE Transactions on Affective Computing, vol. 13, no.2, pp. 992-1004, 1 April-June 2022, doi: 10.1109/TAFFC.2020.2983669

Siddique Latif contributed 75% to this paper. Collectively, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Björn W. Schuller contributed the remainder.

Paper 3:

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Bjorn Wolfgang Schuller. "Self-Supervised Adversarial Domain Adaptation for

Cross-Corpus and Cross-Language Speech Emotion Recognition," in IEEE Transactions on Affective Computing (2022), doi:10.1109/TAFFC.2022.3167013.

Siddique Latif contributed 75% to this paper. Collectively, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller contributed the remainder.

Paper 4:

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller. "Multitask Learning from Augmented Auxiliary Data for Improving Speech Emotion Recognition", in IEEE Transaction on Affective Computing, doi 10.1109/TAFFC.2022.3221749

Siddique Latif contributed 75% to this paper. Collectively, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller contributed the remainder.

Paper 5:

Siddique Latif, Muhammad Asim, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller. "Augmenting generative adversarial networks for speech emotion recognition." In Proceedings of the 2020 Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 521-525. International Speech Communication Association, 2020.

Siddique Latif contributed 70% to this paper. Collectively, Muhammad Asim, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller contributed the remainder.

Paper 6:

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Bjorn W. Schuller. "Deep architecture enhancing robustness to noise, adversarial

attacks, and cross-corpus setting for speech emotion recognition." In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), vol. 4, pp. 2327-2331. International Speech Communication Association (ISCA), 2020

Siddique Latif contributed 75% to this paper. Collectively, Muhammad Asim, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller contributed the remainder.

ACKNOWLEDGEMENTS

I would like to thank my parents, family, and wife for believing in me and supporting me for all those years. I would like to thank to my PhD supervisors: Dr. Rajib Rana, Dr. Ji Zhang, Dr. Sara Khalifa, and Dr. Raja Jurdak. I have learned a lot about audio signal processing and about research in general as well. I also got the chance to improve my presentation and writing skills. I also would like to thank Prof. Julien Epps and Prof. Björn W. Schuller for their help and constructive feedback on my research publications.

This research has been supported by the UniSQ International PhD scholarship and CSIRO-Data61 PhD stipend scholarship.

DEDICATION

To the God Almighty.

To the memory of my parents.

To my family

Who supported me in my life.

To my wife

Who patiently supported me throughout my PhD.

And to my son

may you never stop learning.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1. Background

Speech is a natural mode of communication among humans, and it contains multidimensional information about the intended message, speaker, gender, spoken language, mood, and emotions. Perception of affective states of interlocutors plays a crucial role in the human interpersonal interaction. Researchers are trying to develop methods to enable machines to understand the context of human speech, their languages, and judge their emotions for more natural and harmonious human-computer interaction. Therefore, speech emotion recognition (SER) is becoming an emerging area of research due to its many important applications in healthcare [1,2], forensics sciences [3], smart cars [4], customers service centres [5], and many more.

Human emotions in speech are very complex to model due to their dependency on many factors such as speaker [6], gender [7], culture [8], age [9], dialect [10], and so on. Researchers have been working on different speech features because the performance of SER system heavily dependent upon the choice of these features [11]. For that reason, most of the actual work on SER goes into the designing of pre-processing pipelines or feature extraction or speech transformation techniques [12]. Such feature engineering is crucial but labour-intensive that highlights the weakness of current SER systems. In order to create ease for the applicability of deep learning (DL), it is highly desirable to make learning algorithms less dependent on human ingenuity-based feature engineering techniques. In this way, novel applications would be constructed faster to make real progress towards artificial intelligence (AI).

The recent development in DL algorithms and representation learning has a strong positive impact in speech recognition [13–15]. However, very few studies applied deep representation learning methods in SER [16]. There are still various open research problems in the SER field that can be solved using advanced deep representation learning methods. For instance,

one of the major problems in SER is the unavailability of larger labelled datasets, where supervised learning methods unable to provide the best performance. Here, semi-supervised or self-supervised representation learning methods can be applied to learn from larger unlabelled data and use this knowledge to improve SER performance, where only small size labelled data is available. Similarly, the performance of SER systems drastically drops when they are tested under cross-corpus, cross-language, and noisy conditions. There is a crucial need of designing robust models to enable the emotion identification in unseen conditions. In this regard, deep representation learning techniques can be applied to learn generalised and robust features to solve the performance issue of SER in cross-corpus, cross-lingual, and noisy conditions. In this thesis, I conduct research on the topic of emotional representation learning from speech and develop advanced SER methods to effectively utilise the unlabelled, synthetic, and augmented data in semi-supervised and self-supervised ways.

1.2. Research Aims and Objectives

The aim of this research is to develop new models of emotional representation learning from speech. Specifically, emotional representation learning models are constructed using the advanced deep learning concepts including semi-supervised learning, multi-task learning, and self-supervised learning. This work takes the advantages from the adversarial learning to further enhance the power of these models. Models proposed in this work can be directly employed to learn feature representations from unlabelled data and improve the generalisation and robustness of SER systems. The main objective is divided into the following two main parts:

1. representation learning for improving the performance of SER by utilising the unlabelled, synthetic, and augmented data.
2. representation learning for improving the generalisation and robustness of SER systems against cross-corpus, cross-language, noisy conditions.

To achieve the first objective, this thesis focuses on developing new models to improve SER performance by utilising unlabelled, augmented,

and synthetic data. In objective two, this work explores how generalisation and robustness in SER systems can be achieved using novel representation learning methods.

1.3. Contributions and Outline

In this thesis, I consider improving the performance, generalisation, and robustness of SER systems by utilising the unlabelled, synthetic, and augmented data to improve the performance. Overall, I utilise supervised, semi-supervised, and self-supervised DL methods.

First, I performed literature review on deep representation learning for SER in Chapter 2. I found the SER performance still needs significant improvement and state of-the-art deep representation learning techniques can be utilised to achieve this goal.

In Chapter 3, I propose a multi-task learning technique to effectively utilise the unlabelled data to improve the SER performance. I use gender identifications and speaker recognition as auxiliary tasks, which allow the use of very large datasets, e. g., speaker classification datasets. I show that utilisation of additional data can improve the primary task of SER for which only limited labelled data is available.

In Chapter 4, I propose a self-supervised model that effectively utilise the adversarial learning to improve the generalisation of SER system against cross-corpus and cross language settings. I propose an adversarial dual discriminator (ADDi) network that uses the three-players adversarial game to learn generalised representations without requiring any target data labels. I also introduce a self-supervised ADDi (sADDi) network that utilises self-supervised pre-training with unlabelled data. I propose synthetic data generation as a pretext task in sADDi, enabling the network to produce emotionally discriminative and domain invariant representations and providing complementary synthetic data to augment the system.

In Chapter 5, I present a multi-task framework that effectively utilises the augmented data to improve performance as well as generalisation of the system. I use augmentation type classification and

unsupervised reconstruction as auxiliary tasks. I show that the proposed model can improve the performance as well as the generalisation of the system.

In Chapter 6, I propose a framework that utilises the data augmentation scheme to augment the GAN in feature learning and generation. To show the effectiveness of the proposed framework, this work present results for SER on (i) synthetic feature vectors, (ii) augmentation of the training data with synthetic features, (iii) encoded features in compressed representation. I empirically show that the proposed framework can effectively learn compressed emotional representations as well as it can generate synthetic samples that help improve performance in within-corpus and cross-corpus evaluation.

In Chapter 7, I propose a deeper neural network architecture by fusing dense convolutional network (DenseNet), long short-term memory (LSTM) and highway network to learn powerful discriminative features which are robust to noise. I also propose data augmentation with our network architecture to further improve the robustness. I comprehensively evaluate the architecture coupled with data augmentation against (1) noise, (2) adversarial attacks and (3) cross-corpus settings. Our evaluations show promising results when compared with existing studies and state-of-the-art models.

1.4. Outcomes and Implications

The findings from this work can be utilised in any other audio field: (1) to improve the performance of the system by utilising unlabelled, augmented, synthetic data, (2) to design generalised and robust systems against noisy conditions. The problems being solved here are common across machine learning (ML) communities working on audio, text, and vision domain. Therefore, researchers can utilise the finding from this dissertation to solve the problems of their respective domains. Overall my research finding enrich the deep representation learning techniques and

open a new era of AI where researchers can employ models proposed in this work to solve problems in their field.

CHAPTER 2: PAPER 1 - SURVEY OF DEEP REPRESENTATION LEARNING FOR SPEECH EMOTION RECOGNITION

This chapter presents a comprehensive literature review on the important topic of deep representation learning for SER. It mainly focused on surveying the recent studies related to thesis objectives to find the research gaps for future works. This contribution highlights various techniques, and related challenges, and identifies the important future areas of research.

This paper provides a comprehensive review of deep representation learning from emotional speech and highlights the research gap and future directions around the thesis topic. Based on the findings of the literature review, I attempted to fill the research gaps by developing novel architectures focused on thesis objectives. In the next chapter, I will present a multi-task learning framework that can effectively learn from unlabelled data.

CHAPTER 3: PAPER 2 – MULTI-TASK SEMI-SUPERVISED ADVERSARIAL AUTOENCODING FOR SPEECH EMOTION RECOGNITION

This chapter presents the proposed multi-task learning framework that uses auxiliary tasks for which data is abundantly available. This work is based on objective 1 and focuses on utilising unlabelled data to improve the performance of SER systems. The proposed model shows that utilisation of additional data can improve the primary task of SER for which only limited labelled data is available. It uses gender identification and speaker recognition as auxiliary tasks, which allow the use of very large datasets, e. g., speaker classification datasets. To maximise the benefit of multi-task learning, the proposed model further uses an adversarial autoencoder (AAE) within our framework, which has a strong capability to learn powerful and discriminative features. Furthermore, the unsupervised AAE in combination with the supervised classification networks enables semi-supervised learning which incorporates a discriminative component in the AAE unsupervised training pipeline. This semi-supervised learning essentially helps to improve the generalisation of our framework and thus leads to improvements in SER performance.

This paper explores the idea of utilising the unlabelled data (Objective 1) to improve the performance of SER. Results show that the performance of SER systems can be considerably improved by utilising the unlabelled data for speaker and gender identification in the proposed multi-task learning framework. Next chapter is also based on Objective 1, which present the self-supervised framework that utilise the synthetic data to improve the performance of cross-corpus and cross-language.

CHAPTER 4: PAPER 3 - Self Supervised Adversarial Domain Adaptation for Cross-Corpus and Cross-Language Speech Emotion Recognition

This chapter presents the proposed adversarial dual discriminator (ADDi) network based on both objective 1 and objective 2 to improve SER performance for cross-corpus and cross-language by learning from synthetic data. The proposed model uses the three-players adversarial game to learn generalised representations for cross-corpus and cross-language speech emotion recognition (SER) without requiring any target data labels. I also introduce a self-supervised ADDi (sADDi) network that utilises self-supervised pre-training with unlabelled data. I propose synthetic data generation as a pretext task in sADDi, enabling the network to produce emotionally discriminative and domain invariant representations and providing complementary synthetic data to augment the system. The proposed model is rigorously evaluated using five publicly available datasets in three languages and compared with multiple studies on cross-corpus and cross-language SER. Experimental results demonstrate that the proposed model achieves improved performance compared to the state-of-the-art methods.

This chapter utilised synthetic data (objective 1) in self-supervised learning setting to improve the generalisation of SER system (objective 2). The proposed model able to improve the cross-corpus and cross-language SER by learning from synthetic data. It also helps to minimise the requirement of labelled data by exploiting the potential of synthetic data to enhance SER generalisation. In the next chapter, I explore the idea of utilising augmented data by presenting a novel framework that to improve the generalisation and robustness of SER systems.

CHAPTER 5: PAPER 4 – Multitask Learning from Augmented Auxiliary Data for Improving Speech Emotion Recognition

This chapter is focused on both objectives 1 and 2. It explores the idea of learning from augmented data (objective 1) to improve the generalisation and robustness of SER systems (objective 2). I propose a novel framework that utilise the augmented data in multitask learning (MTL) setting to improve the performance as well the robustness of speech emotion recognition (SER). Recent works in multitask SER require of meta labels for auxiliary tasks, which limits the training of such systems. The proposed model addresses the challenge by proposing a semi-supervised MTL framework (MTL-AUG) that learns generalised representations from augmented data. It utilises the augmentation-type classification and unsupervised reconstruction as auxiliary tasks, which allow training SER systems on augmented data without requiring any meta labels for auxiliary tasks. The semi-supervised nature of MTL-AUG allows for the exploitation of the abundant unlabelled data to further boost the performance of SER. This chapter comprehensively evaluates the proposed framework in the following settings: (1) within corpus, (2) cross-corpus and cross-language, (3) noisy speech, (4) and adversarial attacks. Our evaluations show the improved results compared to existing state-of-the-art methods.

This chapter showed that augmented data can be effectively utilised to improve the generalisation and robustness of SER systems against cross-language, noisy and adversarial attacks. Results are considerably better compared to the state-of-the-art studies and the model can reduce the amount of labelled data by 15-20 %. The next chapter presents the study on utilising the synthetic data to augment the SER system for improvements in performance.

CHAPTER 6: PAPER 5 – AUGMENTING GENERATIVE ADVERSARIAL NETWORKS FOR SPEECH EMOTION RECOGNITION

This chapter focuses on objective 1 and presents the proposed framework that utilises GAN for representation learning and feature generation. In particular, the proposed GAN-based framework is utilised to generate synthetic data to augment the speech emotion classifier. Results show that the proposed framework can effectively learn compressed emotional representations as well as it can generate synthetic samples that help improve performance in within-corpus and cross-corpus evaluation.

This chapter showed that synthetic data generated by the GAN-based framework can be utilised to improve speech emotion classification. Results showed that the synthetic data augmentation helped improve the generalisation of the SER systems, which leads to performance improvement. In the next chapter, I present a novel deep architecture that focuses on objective 2 for robust SER.

CHAPTER 7: PAPER 6 – Deep Architecture Enhancing Robustness to Noise, Adversarial Attacks, and Cross-corpus Setting for Speech Emotion Recognition

This chapter is based on objective 2 and presents a deeper neural network architecture wherein I fuse Dense Convolutional Network (DenseNet), Long short-term memory (LSTM) and Highway Network to learn powerful discriminative features which are robust to noise (Objective 2). This chapter comprehensively evaluates the architecture coupled with data augmentation against (1) noise, (2) adversarial attacks and (3) cross-corpus settings. Evaluations in this chapter on the widely used show promising results when compared with existing studies and state-of-the-art models.

The last chapter focused on deep representation learning for robust speech emotion classification (objective 2). Results showed that deep architectures can learn complex representations that are more robust to noise, cross-corpus, and adversarial attacks. The next chapter will conclude this thesis and highlight the future directions.

CHAPTER 8: DISCUSSION AND CONCLUSION

This thesis presented deep representation learning techniques for developing accurate, generalised, and robust speech emotion recognition (SER) systems. The focus of this study was mainly on the following two objectives:

1. representation learning for improving the performance of SER by utilising the unlabelled, synthetic, and augmented data.

2. representation learning for improving the generalisation and robustness of SER systems against cross-corpus, cross-language, noisy conditions.

This thesis starts with the literature review on deep representation learning, which helped me to find research gaps related to the above-mentioned objectives. Overall, I focused on developing novel deep learning architectures around these objectives and tried to utilise unlabelled, augmented, and synthetic data to improve SER performance, generalisation, and robustness. Next, I conclude my findings on the objective explored in this thesis.

8.1. Representation Learning from Unlabelled Data (Chapter 3)

Chapter 3 focused on objective 1 of learning representation from unlabelled data. Therefore, I propose a semi-supervised adversarial multi-task learning for speech emotion recognition (SER). Specifically, the goal was to put considerable emphasis on a novel technique of utilising unlabelled data for auxiliary tasks through the proposed multi-task semi-supervised learning model to improve the accuracy of the primary task. The model is evaluated on publicly available datasets including IEMOCAP and MSP-IMPROV. Results demonstrated that the purposed performs notably better than (1) the comparable state-of-the-art studies in SER that use similar methodology and/or implementation strategies; (2) supervised single- and multi-task methods based on CNN, and (3) single- and multi-task semi-supervised autoencoders. The proposed approach can overcome the challenge of limited data availability of emotion datasets, which is a significant contribution.

The proposed technique showed that (1) improvement of the auxiliary tasks through the injection of additional unlabelled data predominantly drives the

improvement of the primary task, (2) a combined effort of auxiliary task is better for improving the accuracy of the primary task, than using them individually, (3) for the IEMOCAP and MSP-IMPROV datasets, it is possible to reasonably determine an operating point in terms of how much additional data for the auxiliary task is sufficient, (4) it is important to control the weight of loss function of the unsupervised task in the proposed semi-supervised MTL setting to improve the accuracy of SER, and (5) it is important to control the weight of the loss functions of the primary and secondary tasks to achieve the best possible accuracy for SER.

8.2. Self-supervised Representation Learning (Chapter 4)

Chapter 4 is focused on objective 2 and presented a novel self-supervised learning method that addressed the open challenge of improving the speech emotion recognition (SER) performance in cross-corpus and cross-language settings. I proposed the Adversarial Dual Discriminator (ADDi) network that minimises the domain shift among emotional corpora adversarially. The proposed model focused on exploiting the unlabelled data with self-supervised pre-training and proposed self-supervised ADDi (sADDi). For sADDi, I suggested synthetic data generation as a pretext task, which (1) helped improve the domain generalisation performance of an SER system to tackle the larger domain shift between training and test distributions in cross-corpus and cross-language SER; and (2) produced by product synthetic emotional data to augment the SER system and minimise the requirement of source labelled data. The introduced dual discriminator based ADDi network offers improved cross-corpus and cross-language SER without using any target data labels compared to the single discriminator and other state-of-the-art approaches. This is mainly due to the dual discriminator using a three-players adversarial game to learn generalised representations.

The proposed model achieved considerable improvements in results when partial target labels were fed to the network training. This helped the ADDi to regulate the generalised representations based on the target data by maximally matching the data distributions. The proposed self-supervised pretext task produces synthetic data as a by-product to augment the system to achieve better performance. The proposed model was able to reduce 15-20% source training data using sADDi while achieving similar performance reported by a recent related study [17].

8.3. Representation Learning from Augmented Data (Chapter 5)

Chapter 5 addressed the open challenge of improving the generalisation of speech emotion recognition (SER) with novel auxiliary tasks that do not require any additional labels for training a multi-task learning (MTL) model. This contribution is based on objective 2 and proposed augmentation-type classification and reconstruction as auxiliary tasks that minimise the required labelled data by effectively utilising the information available in the augmented data and facilitating the utilisation of unlabelled data in a semi-supervised way.

The multi-task model offers improved within-corpus, cross-corpus, and cross-language emotion classification. It also shows improved generalisation against noisy speech and adversarial attacks. This is due to the proposed auxiliary tasks that helps the model learn shared representations from augmented data. Considerable improvements in results were found when additional unlabelled data was incorporated into the proposed MTL semi-supervised framework. This helped the model to regulate the generalised representations by learning from unlabelled data. The proposed framework was able to reduce the amount of labelled training data by more than 10% while achieving a similar performance reported by a recent related study [18] using 100% training data.

8.4. Representation Learning from Synthetic Data (Chapter 6)

In chapter 6, I present a novel framework based on a generative adversarial network (GAN) that can learn emotional representation to generate synthetic data. This chapter is based on objective 1 and focused on utilising synthetic data to improve SER performance. I proposed to utilise a data augmentation technique called mixup to augment GAN network in representation learning as well as synthetic feature vector generation. Compared to recent studies, the proposed framework was able to learn better emotional representations in compressed form and to generate synthetic feature vectors that can be effectively utilised to augment the training size of SER for performance improvement.

8.5. Robust Representation Learning (Chapter 7)

Chapter 7 focused on robust representation learning by utilising the deeper architectures. It mainly based on objective 2 and introduced a new hybrid model to build a robust Speech Emotion Recognition (SER) system. The proposed model exploits a Dense Convolutional Network (DenseNet) for

feature extraction, Long Short-Term Memory (LSTM) for contextual learning, and fully connected layers with highway connectivity for discriminative representation learning and produce robust representation. This chapter also proposes data augmentation to further improve the robustness of the architecture. The performance of our proposed technique is evaluated on the widely used IEMOCAP and MSP-IMPROV datasets against noise, adversarial attacks, and cross-corpus settings. Results show that the proposed technique is more robust compared to state-of-the-art models and reveal several valuable information, such as mixup is a better augmentation technique for SER compared to the popular speed perturbation.

Overall, this thesis explored different deep representation learning approaches to improve the performance, generalisation, and robustness of SER systems. The proposed frameworks demonstrate the improved performance compared to the state-of-the-art studied. This thesis showed that deep representation learning techniques have great potential to improve SER performance, generalisation, and robustness by utilising unlabelled, synthetic, and augmented data.

There are multiple future directions for extending the works in this dissertation. The multi-task learning semi-supervised frameworks explored in Chapter 3 and 5 can be benefited by reinforced information. Therefore, it is highly attractive to explore integration of reinforcement learning into such frameworks given a real-life usage such as in a dialogue manager. Researchers can further focus on the tighter coupling between the generation of data and modelling a richer selection of speaker states and traits simultaneously aiming at `holistic' speaker analysis. In addition, it is worth exploring multi-model (text and video) auxiliary tasks to improve the primary task in multi-task learning settings for speech emotion recognition by learning generalised representation.

Self-supervised domain adaptation architectures explored in Chapter 4 of this thesis achieve considerably improved cross-corpus and cross-language SER performance. Future studies can explore these domain adaptive frameworks to model other factors of speech variations, including age, subject, gender, phoneme, noise, and recording device. Further experiments may include evaluating such methods in wild conditions like noisy speech and adversarial noise. It will also be interesting in exploring multimodal pretext task techniques in future self-supervised SER systems. Multimodal human interaction in video and textual form can provide various

opportunities for self-supervised learning to improve cross-corpus and cross-language SER.

REFERENCES

[1] Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. "Speech technology for healthcare: Opportunities, challenges, and state of the art." IEEE Reviews in Biomedical Engineering 14 (2020): 342-356.

[2] Rajib Rana, Siddique Latif, Raj Gururajan, Anthony Gray, Geraldine Mackenzie, Gerald Humphris, and Jeff Dunn. Automated screening for distress: A perspective for the future. European Journal of Cancer Care, page e13033, 2019.

[3] Lisa S Roberts. A forensic phonetic study of the vocal responses of individuals in distress. PhD thesis, University of York, 2012.

[4] Hans-Jörg Vögel, Christian Süß, Thomas Hubregtsen, Elisabeth André, Björn Schuller, Jérôme Härri, Jörg Conradt, Asaf Adi, Alexander Zadorojniy, Jacques Terken, et al. Emotion-awareness for intelligent vehicle assistants: A research agenda. In 2018 IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS), pages 11–15. IEEE, 2018.

[5] Felix Burkhardt, Jitendra Ajmera, Roman Englert, Joachim Stegmann, and Winslow Burleson. Detecting anger in automated voice portal dialogs. In Ninth International Conference on Spoken Language Processing, 2006.

[6] Ni Ding, Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. Speaker variability in emotion recognition-an adaptation based approach. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5101–5104. IEEE, 2012.

[7] Thurid Vogt and Elisabeth André. Improving automatic emotion recognition from speech via gender differentiation. In Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa, 2006.

[8] Siddique Latif, Adnan Qayyum, Muhammad Usman, and Junaid Qadir. Cross lingual speech emotion recognition: Urdu vs. western languages. In 2018 International Conference on Frontiers of Information Technology (FIT), pages 88–93. IEEE, 2018.

[9] Aire Mill, Jüri Allik, Anu Realo, and Raivo Valk. Age-related differences in emotion recognition ability: A cross-sectional study. Emotion, 9(5):619, 2009.

[10] Petri Laukka, Daniel Neiberg, and Hillary Anger Elfenbein. Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations. Emotion, 14(3):445, 2014.

[11] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9):1162–1181, 2006.

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

[13] Siddique Latif, Heriberto Cuayáhuitl, Farrukh Pervez, Fahad Shamshad, Hafiz Shehbaz Ali, and Erik Cambria. "A survey on deep reinforcement learning for audio-based applications." Artificial Intelligence Review (2022): 1-48.

[14] Davis Liang, Zhiheng Huang, and Zachary C Lipton. Learning noise-invariant representations for robust speech recognition. In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 56–63. IEEE, 2018.

[15] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on audio, speech, and language processing, 20(1):30–42, 2011.

[16] Siddique Latif, Rajib Rana, Junaid Qadir, and Julien Epps. Variational autoencoders for learning latent representations of speech emotion: A preliminary study. Proc. Interspeech 2018 (2018): 3107-3111.

[17] Gideon, John, Melvin G. McInnis, and Emily Mower Provost. "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)." IEEE Transactions on Affective Computing 12, no. 4 (2019): 1055-1068.

[18] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Björn Wolfgang Schuller. "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition." IEEE Transactions on Affective computing (2020).

[19] Zhang, Yue, Yifan Liu, Felix Weninger, and Björn Schuller. "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations." In 2017 IEEE International Conference on acoustics, speech, and signal processing (ICASSP), pp. 4990-4994. IEEE, 2017.