

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.



MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA

UNIVERSITY OF NEW ZEALAND

Deep Learning for Speech Enhancement

A thesis presented in partial fulfilment of the
requirements for the degree of

Doctor of Philosophy
in
Computer Science

at Massey University, Albany,
New Zealand.

Yuanhang Qiu

2022

Abstract

Speech enhancement, aiming at improving the intelligibility and overall perceptual quality of a contaminated speech signal, is an effective way to improve speech communications. In this thesis, we propose three novel deep learning methods to improve speech enhancement performance.

Firstly, we propose an adversarial latent representation learning for latent space exploration of generative adversarial network based speech enhancement. Based on adversarial feature learning, this method employs an extra encoder to learn an inverse mapping from the generated data distribution to the latent space. The encoder establishes an inner connection with the generator and contributes to latent information learning.

Secondly, we propose an adversarial multi-task learning with inverse mappings method for effective speech representation. This speech enhancement method focuses on enhancing the generator's capability of speech information capture and representation learning. To implement this method, two extra networks are developed to learn the inverse mappings from the generated distribution to the input data domains.

Thirdly, we propose a self-supervised learning based phone-fortified method to improve specific speech characteristics learning for speech enhancement. This method explicitly imports phonetic characteristics into a deep complex convolutional network via a contrastive predictive coding model pre-trained with self-supervised learning.

The experimental results demonstrate that the proposed methods outperform previous speech enhancement methods and achieve state-of-the-art performance in terms of speech intelligibility and overall perceptual quality.

Acknowledgements

I would like to express my sincere gratitude to all the people who have supported me on my PhD journey.

First and foremost, I am extremely grateful to my main supervisor, Professor Ruili Wang, my co-supervisors, Dr Feng Hou and Professor Johan Potgieter, for their invaluable advice, continuous support, and endless patience during my PhD study. They encouraged me in my academic research and daily life. Their insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. Without their valuable support, comments, suggestions, and persistent encouragement, it would be impossible for me to complete my PhD study.

I would also like to thank my friends, lab mates, and colleagues in Professor Wang's research team for a cherished time spent together in the lab and their friendships, encouragement and valuable suggestions.

I am thankful to many faculty members in the School of Natural and Computational Sciences. They provided valuable guidance and support through my doctoral research.

I greatly acknowledge the financial support from the China Scholarship Council and the Catalyst: Strategic – New Zealand-Singapore Data Science Research Programme towards my study and research.

My appreciation also goes out to my family for their encouragement and support all through my studies. You are always there for me. That will support me to move forward through my whole life.

Publications

The following research papers have been published in or submitted to an international journal and conferences:

- **Yuanhang Qiu** and Ruili Wang*. Adversarial Latent Representation Learning for Speech Enhancement. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH), Virtual Event, Shanghai, China, pages 2662 - 2666, 2020. CORE rank A. (Refer to Chapter 2)
- **Yuanhang Qiu**, Ruili Wang*, Feng Hou, Satwinder Singh, Zhizhong Ma and Xiaoyun Jia. Adversarial Multi-task Learning with Inverse Mapping for Speech Enhancement. Under review of Applied Soft Computing. JCR Q1. (Refer to Chapter 3)
- **Yuanhang Qiu**, Ruili Wang*, Satwinder Singh, Zhizhong Ma and Feng Hou. Self-Supervised Learning Based Phone-Fortified Speech Enhancement. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH), Hybrid Event, Brno, Czechia, pages 211 - 215, 2021. CORE rank A. (Refer to Chapter 4)
- Satwinder Singh, Ruili Wang* and **Yuanhang Qiu**. DEEPF0: End-To-End Fundamental Frequency Estimation for Music and Speech Signals. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Event, Toronto, Ontario, Canada, pages 61 - 65, 2021. CORE rank B

Contents

Abstract	i
Acknowledgements	ii
Publications	iii
List of Figures	ix
List of Tables	xiii
1 Introduction	1
1.1 Overview of Speech Enhancement	1
1.2 Motivations	3
1.3 Contributions	4
1.4 Organization of Thesis	5
2 Adversarial Latent Representation Learning for Speech Enhancement	13
2.1 Introduction	14
2.2 Related Work	16
2.2.1 GAN-based Speech Enhancement	16
2.2.2 Latent Space	19

2.3	Adversarial Latent Representation Learning	20
2.3.1	ALRL	20
2.3.2	Multi-head Self-attention	22
2.4	Experiments	23
2.4.1	Database	23
2.4.2	Setup	24
2.5	Results	24
2.6	Conclusions	27
3	Adversarial Multi-Task Learning with Inverse Mappings for Speech Enhancement	33
3.1	Introduction	34
3.2	Related Work	38
3.2.1	Speech Enhancement with Adversarial Networks	38
3.2.2	Inverse Mapping Learning	40
3.3	Methodology	42
3.3.1	Loss Functions	42
3.3.2	Model Architecture Setup	45
3.4	Experiments	48
3.4.1	Database	48
3.4.2	Setup	50
3.4.3	Evaluation Metrics	53
3.5	Results and Discussions	53
3.6	Conclusions	60
4	Self-Supervised Learning Based Phone-Fortified Speech Enhancement	71
4.1	Introduction	72
4.2	Related Work	74

4.2.1	Phase Information Preservation	74
4.2.2	Speech Representation Learning	75
4.2.2.1	Wav2vec	75
4.2.2.2	Vq-wav2vec	77
4.3	The Proposed Method	77
4.4	Experiments	80
4.4.1	Database	80
4.4.2	Setup	80
4.4.3	Evaluation Metrics	82
4.5	Results	82
4.6	Conclusions	84
5	Summary	93
5.1	Research Summary	93
5.2	Future Work	95
A	Statement of Contribution	99

List of Figures

2.1	The basic framework of GAN-based speech enhancement. The Generator (G) converts noisy signal and latent vector to the generated samples. The Discriminator (D) is trained to distinguish the generated samples and clean speech as fake or real.	15
2.2	The framework of adversarial latent representation learning. The Encoder (E) processes the generated samples and provides latent information for model training.	21
3.1	The framework of GAN-based speech enhancement. The Generator (G) consumes raw noisy speech and latent vector (i.e., random noise) as input. The Discriminator (D) is a binary classifier aiming to judge the similarity between the generated sample and raw clean speech. . .	36
3.2	The framework of our proposed method. Based on basic GAN architecture, the Generator (G) receives raw noisy speech and random noise as input. The Discriminator (D) gives the judgment (i.e., fake or real) of the generated sample and raw clean speech. The networks P and Q are proposed to establish the inverse mappings from the generated distribution to the input data domain for information capture and representation learning.	44

3.3	The details of Generator (G). The downsampling adopts 2D convolutional kernels followed by PReLU for information capture. The upsampling adopts 2D transposed convolutional kernels followed by PReLU for sample reconstruction. Latent vector z gets concatenated with the condensed representation of the bottleneck layer. The skip connections are used to boost the stability of model training.	46
3.4	The details of Discriminator (D). D is a binary classifier to give the judgment (fake or real) of the ground truth and the generated sample. The main components are the 2D convolutional kernels followed by Virtual Batch Normalization (VBN) and LeakyReLU for distinguishable information learning.	47
3.5	Spectrograms of selected utterance (SNR=2.5dB) enhanced with our method.	54
3.6	Spectrograms of selected utterance (SNR=7.5dB) enhanced with our method.	55
3.7	Spectrograms of selected utterance (SNR=12.5dB) enhanced with our method.	56
3.8	Spectrograms of selected utterance (SNR=17.5dB) enhanced with our method.	57
4.1	The framework of wav2vec and vq-wav2vec. For wav2vec, the Encoder (E) network maps the raw audio to a dense representation z . z is aggregated into a Context (C) network for representation c , which refers to the contrastive loss calculation (l) with the future samples. With the addition of a quantized (q) layer, this framework represents vq-wav2vec.	76

4.2	The framework of our proposed speech enhancement model based on a complex U-Net and CPC. The complex U-Net estimates a complex-valued ratio mask with the fused noisy speech representation. The mask can filter the noisy spectrum and achieve the enhanced spectrum with inverse STFT. The CPC-based pre-trained model extracts linguistic information of an enhanced waveform and paired clean waveform for PFP loss calculation.	79
-----	--	----

List of Tables

2.1	The evaluation results of different methods in quality and intelligibility. The results include four SNR conditions (i.e. 17.5 dB, 12.5dB, 7.5dB, and 2.5 dB) and the overall values. "†" denotes that we reproduced the results with the provided open resources. The best scores are highlighted in bold.	26
3.1	The evaluation results of our method compared with previous methods including Wiener filtering [30], SEGAN [41], SERGAN [2], MMSE-GAN [48], BiLSTM [14], CRN-MSN [52], NAAGN [8]. All the presented methods were trained with the 28-speaker database. "†" denotes that we reproduced the results with the provided open resource. "-" denotes that the result is not reported or not available. The best scores are highlighted in bold.	49
3.2	The unfolded evaluation results on different SNR values (i.e., 17.5 dB, 12.5dB, 7.5dB, 2.5dB, and overall). We evaluate SEGAN [41], SERGAN [2], ARL [44] and our method with more comprehensive speech quality and intelligibility metrics on the 28-speaker database. "†" denotes that we reproduced the results with the provided open resources.	51
3.3	The evaluation results of SEGAN [41], SERGAN [2] and our method with different scales of training database in terms of speech quality and intelligibility. "†" denotes that we reproduced the results with the provided open resources.	52

-
- 4.1 The evaluation results of various methods with the 28-speaker VCTK training data. The compared methods are Wiener filtering [17], ALRL [24], BiLSTM [7], CRN-MSN [34], AMTL-IM [25], NAAGN [5], U-NetC [2, 37], PHASEN [44], HiFi-GAN [31], T-GSA [14]. “-” denotes that the result is not reported or not available. “†” denotes that we reproduced the results with the provided open resource. The best scores are highlighted in bold. 81
- 4.2 The evaluation results of $W2V_{FSA}$ with 56-speaker and 84-speaker (the mixture of 28-speaker and 56-speaker). Meanwhile, all scenes with different SNR and noise types are demonstrated separately. The best scores are highlighted in bold. 83

Chapter 1

Introduction

This chapter provides an overview of this thesis. The background of speech enhancement is introduced briefly in Section 1.1, where the issues with current speech enhancement approaches are analyzed. The motivations are explained in Section 1.2. The contributions of this thesis are summarized in Section 1.3. Finally, the organization of this thesis is listed in Section 1.4.

1.1 Overview of Speech Enhancement

Speech is the most common form of human-to-human communication [18]. Natural human speech can deliver a lot of information including basic context meaning, speakers' present status information, such as identity, emotion, gender, age bracket, etc., to receivers [14]. However, various disturbing noises usually contaminate the natural speech signals and compromise the effectiveness of information delivery in real life [21]. Therefore, there has been a focused research topic to reduce the negative effects of disturbing noises and improve speech information delivery in speech signal processing currently.

Speech enhancement, aiming to improve the intelligibility and overall perceptual quality of a contaminated speech signal, is one of the most important speech signal processing technologies [2, 3]. The intelligibility is a measurement of how comprehensible a speech signal is, while the perceptual quality measures how easy it is for a listener to perceive the content of a speech signal. Normally, a perceptual high-quality speech sounds more natural, rhythmic, yet less raspy, hoarse or scratchy, etc. In practice, speech enhancement is widely used in many applications such as mobile communications [3], hearing aids [5, 19], and robust speech recognition [33].

Generally, the disturbing noises can be categorized into stationary additive noise (e.g., fan noise) and non-stationary convolutional noise (e.g., room reverberation), which can badly degrade both the intelligibility and perceptual quality of speech signals [39]. Correspondingly, many kinds of speech enhancement approaches, such as classic digital signal processing [4, 8], traditional machine learning [13, 15, 29, 34], and novel deep neural networks [10, 12, 22, 31, 36, 37, 40, 41], were proposed to suppress the stationary or non-stationary noises and improve enhancement performance.

The classic digital signal processing approaches (e.g., spectral subtraction [4], Wiener filtering [21], minimum mean square error [8]) usually manipulate spectrum magnitudes of noisy speech signals and perform well in additive noise suppression. However, due to the sophisticated statistical properties of interactions between speech and noise signals [37], the digital signal processing approaches often result in speech distortion and residual noise to some extent, in a low signal-to-noise ratio scenario. Moreover, the digital signal processing approaches also struggle to deal with non-stationary noises in complex scenarios.

Besides, a series of machine learning approaches (e.g., Gaussian mixture model [15], hidden Markov model [34], and non-negative matrix factorisation [13]) were also

applied to speech enhancement. Those approaches, based on data distribution modelling and analysis, mitigate speech distortion and residual noise problems and improve speech enhancement greatly. However, learning effective speech representation and improving non-stationary speech enhancement in more noisy acoustic environments are still long-standing and challenging tasks.

With the progressive development of artificial intelligence algorithms, deep learning has shown its revolutionary capability in many research areas, such as computer vision [16, 35], natural language processing [7], and recommender system [38]. As one of the most important research topics of speech signal processing, speech enhancement has also applied deep learning based models such as the denoising autoencoder [22, 30], Recurrent Neural Networks (RNN) [12, 32], Convolutional Neural Networks (CNN) [9, 24], and Generative Adversarial Networks (GAN) [1, 11, 25] to improve speech representation learning and enhancement performance. With powerful learning and inference capability, data-driven deep learning is suitable for complex speech signal processing. Based on that, this thesis focuses on deep learning based speech enhancement. The related studies are presented in detail in the following chapters.

1.2 Motivations

Although deep learning based speech enhancement methods have greatly alleviated the existing problems that are difficult to solve using the traditional methods, there are still several research points to be considered:

- **Effective speech representation learning.** Current deep learning based methods require large amounts of data for model training. Learning effective data representation from the large amounts of speech data is the key

to obtain outstanding performance [6]. For speech enhancement, learning effective speech representation and exploring sufficient latent information with advanced model can significantly improve enhancement performance [23], however, are still huge challenges in current speech signal processing.

- **Appropriate loss functions design.** For GAN-based speech enhancement, improving the robustness and effectiveness of model training is one of the most challenging aspects. An appropriate loss function with empirical hyperparameters setup can greatly stabilize model training [20] and further improve the performance of speech enhancement.
- **Specific speech information preservation.** Previous traditional speech enhancement methods usually ignore the specific speech information such as phase information and phonetic characteristics, which have been proven to be effective for mitigating residual noise and improving speech signal reconstruction [17]. Thus, explicitly learning speech phase information and utilizing phonetic characteristics is an effective way of improving speech enhancement.

1.3 Contributions

To address the issues mentioned above and improve speech enhancement, three novel methods are proposed in this thesis, summarized below:

- For sufficient latent space exploration, we propose a novel Adversarial Latent Representation Learning (ALRL) method for speech enhancement [26]. Based on adversarial feature learning, ALRL employs an extra encoder to learn an inverse mapping from the generated data distribution to the latent space. The encoder establishes an inner connection with the generator and facilitates relevant latent information learning. A new loss function is proposed to implement

the encoder mapping. In addition, the multi-head self-attention is also applied to learn the long-range dependencies of speech utterances.

- For effective speech representation learning, we propose a novel adversarial multi-task learning with inverse mappings method for speech enhancement [27]. This method focuses on enhancing the generator’s capability of speech information capturing and representation learning. To implement this method, two extra networks (namely P and Q) are developed to establish the inverse mappings from the generated distribution to the input data domains. Correspondingly, the latent loss and equilibrium loss are proposed for the inverse mappings learning and the enhancement model training based on the original adversarial loss.
- For specific speech characteristics preservation, we propose a novel Self-Supervised learning based Phone-Fortified (SSPF) method for speech enhancement [28]. This method explicitly incorporates phonetic characteristics into a deep complex convolutional network via a Contrastive Predictive Coding (CPC) model pre-trained with self-supervised learning. The deep complex network can effectively deal with complex-valued spectrums of speech signals and greatly improve speech phase information learning.

1.4 Organization of Thesis

This is a thesis with publication, which is organised in the following ways:

- (i) Literature review corresponding to each proposed method is presented in each chapter; (ii) All references related to each chapter are listed at the end of each chapter.**

Chapter 2 presents the proposed adversarial latent representation learning method for latent information exploration and adversarial feature learning.

Chapter 3 presents the proposed adversarial multi-task learning method for speech information capture and representation learning.

Chapter 4 presents the proposed self-supervised learning based phone-fortified method for phonetic characteristics extracting and speech enhancement improvement.

Chapter 5 summarizes this thesis and discusses the future work.

References

- [1] Deepak Baby and Sarah Verhulst. SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 106–110. IEEE, 2019. doi: <https://doi.org/10.1109/ICASSP.2019.8683799>.
- [2] Jacob Benesty, Shoji Makino, and Jingdong Chen. *Speech enhancement*. Springer Science & Business Media, 2006. doi: <https://doi.org/10.1007/3-540-27489-8>.
- [3] Jacob Benesty, Jesper Rindom Jensen, Mads Graesboll Christensen, and Jingdong Chen. *Speech enhancement: A signal subspace perspective*. Academic Press, 1 edition, 2014. doi: <https://doi.org/10.1016/B978-0-12-800139-4.00009-8>.
- [4] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, 27(2):113–120, 1979. doi: <https://doi.org/10.1109/TASSP.1979.1163209>.

-
- [5] Komal R Borisagar, Rohit M Thanki, and Bhavin S Sedani. *Speech enhancement techniques for digital hearing aids*. Springer, 2019. doi: <https://doi.org/10.1007/978-3-319-96821-6>.
- [6] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2172–2180, 2016.
- [7] Li Deng and Yang Liu. *Deep learning in natural language processing*. Springer, 2018. doi: <https://doi.org/10.1007/978-981-10-5209-5>.
- [8] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, 32(6):1109–1121, 1984. doi: <https://doi.org/10.1109/TASSP.1984.1164453>.
- [9] Ori Ernst, Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger. Speech dereverberation using fully convolutional networks. In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, pages 390–394. IEEE, 2018. doi: <https://doi.org/10.23919/EUSIPCO.2018.8553141>.
- [10] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 26(9):1570–1584, 2018. doi: <https://doi.org/10.1109/TASLP.2018.2821903>.
- [11] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2031–2041. PMLR, 2019.
- [12] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee. Densely connected progressive learning for LSTM-based speech enhancement. In *Proceedings of the*

- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5054–5058. IEEE, 2018. doi: <http://doi.org/10.1109/ICASSP.2018.8461861>.
- [13] Jürgen T Geiger, Jort F Gemmeke, Björn Schuller, and Gerhard Rigoll. Investigating NMF speech enhancement for neural network based acoustic models. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.
- [14] Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011. doi: <https://doi.org/10.1121/1.4742973>.
- [15] Jiucang Hao, Te-Won Lee, and Terrence J Sejnowski. Speech enhancement using gaussian scale mixture models. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 18(6):1127–1136, 2009. doi: <https://doi.org/10.1109/TASL.2009.2030012>.
- [16] Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. Deep learning vs. traditional computer vision. In *Proceedings of the Computer Vision Conference (CVC)*, volume 943, page 128. Springer, 2019. doi: https://doi.org/10.1007/978-3-030-17795-9_10.
- [17] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2472–2476, 2020. doi: <https://doi.org/10.21437/Interspeech.2020-2537>.
- [18] AR Jayan. *Speech and Audio Signal Processing*. PHI Learning Pvt. Ltd., 2017.
- [19] Mathew Shaji Kavalekalam, Jesper Kjar Nielsen, Jesper Bunsow Boldt, and Mads Grasboll Christensen. Model-based speech enhancement for intelligibility improvement in binaural hearing aids. *IEEE/ACM Transactions on Audio,*

- Speech, and Language Processing (TASLP)*, 27(1):99–113, 2019. doi: <https://doi.org/10.1109/TASLP.2018.2872128>.
- [20] Morten Kolbæk, Zheng-Hua Tan, Søren Holdt Jensen, and Jesper Jensen. On loss functions for supervised monaural time-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 28:825–838, 2020. doi: <https://doi.org/10.1109/TASLP.2020.2968738>.
- [21] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. 2013. ISBN 1466504218. doi: <https://doi.org/10.1201/9781420015836>.
- [22] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, volume 2013, pages 436–440, 2013.
- [23] Zhiheng Ouyang, Hongjiang Yu, Wei-Ping Zhu, and Benoit Champagne. A fully convolutional neural network for complex spectrogram processing in speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5756–5760. IEEE, 2019. doi: <http://doi.org/10.1109/ICASSP.2019.8683423>.
- [24] Se Rim Park and Jin Won Lee. A fully convolutional neural network for speech enhancement. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1993–1997, 2017. doi: <https://doi.org/10.23919/10.21437/Interspeech.2017-1465>.
- [25] Santiago Pascual, Antonio Bonafonte, and Joan Serra. SEGAN: speech enhancement generative adversarial network. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3642–3646, 2017. doi: <https://doi.org/10.21437/Interspeech.2017-1428>.
- [26] Yuanhang Qiu and Ruili Wang. Adversarial latent representation learning for

- speech enhancement. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2662 – 2666, 2020. doi: <https://doi.org/10.21437/Interspeech.2020-1593>.
- [27] Yuanhang Qiu, Ruili Wang, Feng Hou, Satwinder Singh, Zhizhong Ma, and Xiaoyun jia. Adversarial multi-task learning with inverse mapping for speech enhancement. *Under review of Applied Soft Computing*, 2021.
- [28] Yuanhang Qiu, Ruili Wang, Satwinder Singh, Zhizhong Ma, and Feng Hou. Self-supervised learning based phone-fortified speech enhancement. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 211 – 215, 2021. doi: <https://doi.org/10.21437/Interspeech.2021-734>.
- [29] Björn Schuller, Felix Weninger, Martin Wöllmer, Yang Sun, and Gerhard Rigoll. Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4562–4565. IEEE, 2010. doi: <https://doi.org/10.1109/ICASSP.2010.5495567>.
- [30] Prashanth Gurunath Shivakumar and Panayiotis G Georgiou. Perception optimized deep denoising autoencoders for speech enhancement. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3743–3747, 2016. doi: <https://doi.org/10.21437/Interspeech.2016-1284>.
- [31] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. HiFi-GAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks. In *Proceedings of the 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4506–4510, 2020. doi: <http://doi.org/10.21437/Interspeech.2020-2143>.

-
- [32] Lei Sun, Jun Du, Li-Rong Dai, and Chin-Hui Lee. Multiple-target deep learning for LSTM-RNN based speech enhancement. In *Proceedings of the Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pages 136–140. IEEE, 2017. doi: <https://doi.org/10.1109/HSCMA.2017.7895577>.
- [33] Yan-Hui Tu, Jun Du, and Chin-Hui Lee. Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(12):2080–2091, 2019. doi: <https://doi.org/10.1109/TASLP.2019.2940662>.
- [34] Hadi Veisi and Hossein Sameti. Speech enhancement using hidden markov models in mel-frequency domain. *Speech Communication*, 55(2):205–220, 2013. doi: <https://doi.org/10.1016/j.specom.2012.08.005>.
- [35] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018. doi: <https://doi.org/10.1155/2018/7068349>.
- [36] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1):65–68, 2013. doi: <https://doi.org/10.1109/LSP.2013.2291240>.
- [37] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 23(1):7–19, 2014. doi: <https://doi.org/10.1109/TASLP.2014.2364452>.
- [38] Yonghong Yu, Yang Gao, Hao Wang, and Ruili Wang. Joint user knowledge and matrix factorization for recommender systems. *World Wide Web*, 21(4): 1141–1163, 2018. doi: <https://doi.org/10.1007/s11280-017-0476-7>.
- [39] Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. Deep learning for environmentally robust

- speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–28, 2018. doi: <https://doi.org/10.1145/3178115>.
- [40] Yan Zhao, DeLiang Wang, Ivo Merks, and Tao Zhang. DNN-based enhancement of noisy and reverberant speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE, 2016. doi: <https://doi.org/10.1109/ICASSP.2016.7472934>.
- [41] Yan Zhao, Buye Xu, Ritwik Giri, and Tao Zhang. Perceptually guided speech enhancement using deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5074–5078. IEEE, 2018. doi: <https://doi.org/10.1109/ICASSP.2018.8462593>.

Chapter 2

Adversarial Latent Representation Learning for Speech Enhancement

In this chapter, we propose a novel Adversarial Latent Representation Learning (ALRL) method for speech enhancement. Based on adversarial feature learning, ALRL employs an extra encoder to learn an inverse mapping from the generated data distribution to the latent space. The encoder establishes an inner connection with the generator and provides relevant latent information for adversarial feature modelling. A new loss function is proposed to implement the encoder mapping. In addition, the multi-head self-attention is also applied to the encoder for learning of long-range dependencies and further effective adversarial representations. The experimental results demonstrate that ALRL outperforms current GAN-based speech enhancement methods.

2.1 Introduction

Speech enhancement aims to improve the intelligibility and overall perceptual quality of contaminated speech signals [19]. There are many practical applications such as telephone communications [20], hearing-aid devices [17], and human-computer interactions [31], which regard speech enhancement as an essential operation for different purposes and processing stages. More complicated and critical application scenarios require higher performance of speech enhancement.

Classic digital signal processing methods of speech enhancement (e.g. Wiener filtering [33], spectral subtraction [4]) perform well in specific additive noise suppressing. However, these methods are difficult to process assorted unknown noise interference satisfactorily. To solve this problem, learning appropriate representation of noise data distribution is a key procedure in current data-driven approaches.

Recently, deep learning based methods have shown revolutionary information learning and reconstruction property in many research areas. Profiting from this, a series of neural network based speech enhancement methods such as denoising autoencoder [6, 29], Long Short-Term Memory (LSTM) based [11], Convolutional Neural Networks (CNN) based methods [14, 24] were also developed for improving speech enhancement performance. Particularly, the Generative Adversarial Networks (GAN) [12, 26, 27], which was originally proposed with artful architecture design for high-quality images generation in computer vision, has been applied successfully to speech enhancement [18, 25].

GAN consists of a generator and a discriminator, which are trained adversarially up to the Nash Equilibrium [12]. For speech enhancement as shown in Figure 2.1, the generator usually takes in noisy speech and extra noise distributions (i.e. latent vectors) as input and exports targeted data distribution. The discriminator is considered as a classifier trained to distinguish generated samples and clean speech

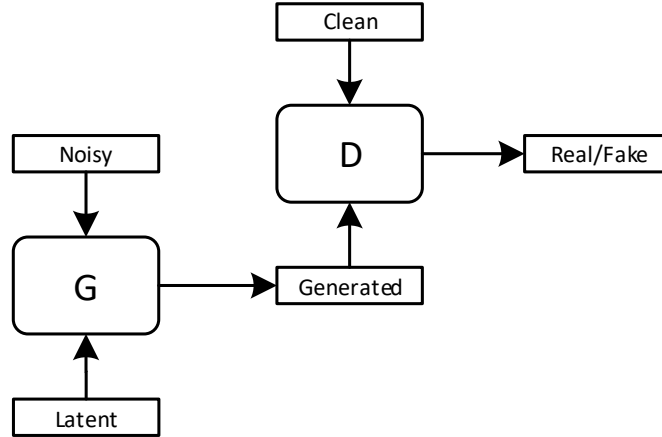


FIGURE 2.1: The basic framework of GAN-based speech enhancement. The Generator (G) converts noisy signal and latent vector to the generated samples. The Discriminator (D) is trained to distinguish the generated samples and clean speech as fake or real.

as fake or real. With effective representation learning and enhancement performance, GAN-based speech enhancement methods have attracted a good proportion of attention[2, 25] in speech enhancement.

However, no attention has been paid to latent space for representation learning in speech enhancement. Initially, GAN can generate high-quality targets from latent vectors based on real data distribution. Thus, in speech enhancement, we hypothesise that the latent vectors play an important part in representation learning conditional upon explicit noisy data distribution.

In this work, we propose a novel GAN-based method named Adversarial Latent Representation Learning (ALRL), which employs an extra encoder inversely mapping the generated data distribution to the latent space, for speech enhancement improvement. In particular, the encoder attempts to build an inner connection with the generator and provides relevant representation information for the modelling of the adversarial features. The new architecture remodels the inner projection from the concatenated input of noisy speech and latent vectors to the clean speech distribution. To implement the encoder mapping, we propose a new encoder mapping loss

function, which captures latent representation by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors. Also, we combine the encoder loss with the relativistic loss [16] to further improve the effectiveness of information learning between the generator and discriminator. In the meanwhile, the multi-head self-attention mechanism [32] is also applied to the encoder in our ALRL for long-range dependencies capturing and further effective representation learning.

This chapter is organized as follows. The related work is given in Section 2.2. The details of our adversarial latent representation learning are introduced in Section 2.3. Sections 2.4 gives the design of the experiments, and the experimental results are presented in Section 2.5. Finally, the conclusions and future work are shown in Section 2.6.

2.2 Related Work

In this section, we introduce related GAN-based speech enhancement methods and present a preliminary investigation of latent space in GAN-based architecture. Based on these works, our ALRL is proposed to learn semantic representation and improve speech enhancement performance.

2.2.1 GAN-based Speech Enhancement

Recently, the GAN-based models have derived huge progress on semantic representation learning and improved speech enhancement performance significantly. Speech Enhancement GAN (SEGAN) is one of the most famous frameworks proposed for

time-domain speech enhancement with improved conditional GAN [25], which combined the conditional GAN with the Least-Squares GAN (LSGAN) together to further alleviate vanishing gradients. This modification is proved to be effective in performance improvement. Below is the loss function of its discriminator:

$$L_D = \frac{1}{2} E_{x \sim P_x, x_c \sim P_{x_c}} [(D(x, x_c) - 1)^2] + \frac{1}{2} E_{z \sim P_z, x_c \sim P_{x_c}} [(D(G(z, x_c), x_c))^2] \quad (2.1)$$

and its generator:

$$L_G = \frac{1}{2} E_{z \sim P_z, x_c \sim P_{x_c}} [(D(G(z, x_c), x_c) - 1)^2] \quad (2.2)$$

where x_c denotes noisy speech; x denotes clean speech; and z denotes random noise distribution (i.e. latent information).

SEGAN operated on raw speech waveform directly rather than the processed spectral features, which is considered to be able to preserve original sequential information such as phase information effectively. SEGAN worked end-to-end and was trained adversarially based on GAN. The fully convolutional architecture consists of downsampling and upsampling modules (i.e. encoder and decoder). The random noise z (i.e. latent vectors) was added to the bottleneck layer for information compensation whereas without the further introduction of it. As we know, this work applied conditional GAN to speech enhancement firstly and obtained outstanding performance. In the meanwhile, conditional GANs were also applied to noise-robust speaker verification [23] and speech recognition [7]. However, the latent space still has not been explored thoroughly in speech signal processing.

Speech Enhancement Relativistic GAN (SERGAN) [2] is another framework exploring speech enhancement based on GAN. In the standard GAN, the discriminator is

developed to estimate the probability that the original data is real and the generated data is fake, on the contrary, the generator is trained to increase the probability that fake data is real. However, it should simultaneously decrease the probability that real data is real when the generator learns to increase that probability. To accomplish the assumption, the relativistic GAN [16] was proposed for a more stable model and higher quality data samples. Below is the loss function of the discriminator:

$$L_D = -E_{x \sim P_x, x_c \sim P_{x_c}} [\log(\sigma(C(x, x_c) - C(G(z, x_c), x_c)))] \quad (2.3)$$

and generator:

$$L_G = -E_{x \sim P_x, x_c \sim P_{x_c}} [\log(\sigma(C(G(z, x_c), x_c) - C(x, x_c)))] \quad (2.4)$$

where σ is the sigmoid non-linearity, and $C(x)$ denotes the discriminator without the final sigmoid layer. $D(x) = \sigma(C(x))$.

SERGAN built a closer information connection between the generator and discriminator for speech enhancement. Moreover, the gradient penalty was also utilized to stabilize model training and improve enhancement performance. The method held a similar architecture with SEGAN but adopted a new loss function to boost information communication between the generator and discriminator. However, this work still did not explore latent space further.

In addition, the improved Speech Enhancement GAN (iSEGAN) [1] conducted a preliminary experiment to explore the impact of the latent vectors on a speech enhancement model by comparing the performance of the model trained with and without latent vector. The results showed that the latent vectors could slightly affect the model performance but were helpful to stabilize model training.

These methods mentioned above attempt to obtain performance gains by modifying

model architecture. Also, some works explore data space for better model performance.

2.2.2 Latent Space

The latent vectors may be used by the generator in a highly entangled way, causing the individual dimensions of latent vectors to not correspond to semantic features of the data. For image generation, Chen et al. [5] proposed to adopt a mutual information strategy for inducing latent vectors. The method decomposed the input noise vectors into a set of semantically meaning factors of variation rather than using single unstructured noise vectors. The work discovered that these latent factors can target salient semantic features of data distribution.

Similarly, Donahue et al. [8] noticed that GAN models could capture semantic variation from latent space, however, have no means of projecting data back into the latent space. This resulted in the architecture ignoring much of the useful information presenting in the structure of the data itself. In addition, interpolations in the latent space of the generator produced smooth and plausible semantic variations and made the model learn to associate particular latent directions with specific features. Thus, the Bidirectional Generative Adversarial Networks (BiGAN) was proposed to learn a generative mapping from simple latent distributions to arbitrarily complex data distribution [8]. Another similar work about latent space exploration was proposed to map training examples in the data space to the space of latent variables as well [10].

Inspired by the mentioned work, we infer that the latent space plays an important role in a generative model for semantic representation capturing. Thus, we propose an adversarial latent representation learning method for speech enhancement.

2.3 Adversarial Latent Representation Learning

In this section, we introduce the details of our Adversarial Latent Representation Learning (ALRL) method for speech enhancement.

2.3.1 ALRL

The related studies [5, 8] show that the latent space can target salient semantic features of data distribution and provide effective guidance for information learning. One effective method is to project the data back into the latent space. In our work, an encoder is built for latent representation learning whereas our encoder will be trained for inverse mapping from generated samples to latent space as shown in Figure 2.2. An inner connection is established by the encoder to learn more useful information for speech representation learning. To implement the encoder mapping, we propose a new encoder loss function, which captures latent representation by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors.

To further improve the effectiveness of information learning, we combine the encoder loss with the relativistic loss function [16]. The encoder, generator and discriminator will be trained simultaneously. Below is our new loss function for the generator:

$$\begin{aligned}
 L_G = & - E_{x \sim P_x, x_c \sim P_{x_c}} [\log(\sigma(C(G(z, x_c), x_c) - C(x, x_c)))] \\
 & - E_{z \sim P_z, x_c \sim P_{x_c}} [\|E(G(z, x_c)) - z\|_2^2]
 \end{aligned} \tag{2.5}$$

where E is defined by calculating the squared Euclidean [9] loss. The new loss function improves the semantic representation learning of the generator. To further avoid vanishing gradients, the gradient penalty regularization is also used in discriminator as proposed in [2]. Below is the discriminator:

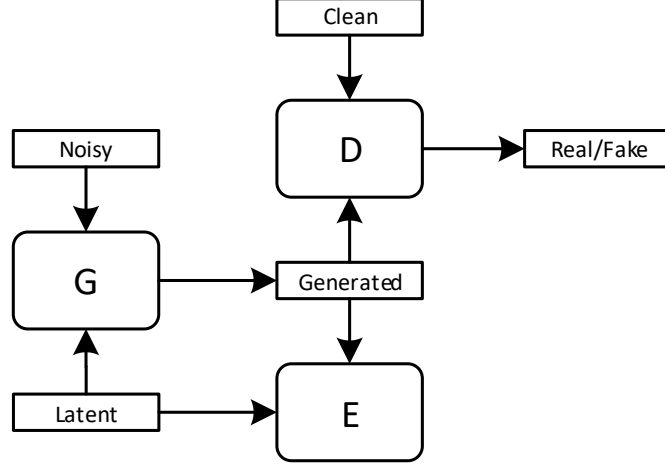


FIGURE 2.2: The framework of adversarial latent representation learning. The Encoder (E) processes the generated samples and provides latent information for model training.

$$\begin{aligned}
 L_D = & - E_{x \sim P_x, x_c \sim P_{x_c}} [\log(\sigma(C(x, x_c) - C(G(z, x_c), x_c)))] \\
 & - \lambda E_{\tilde{x}, x \sim P_{(\tilde{x}, x)}} [(\|\nabla_{\tilde{x}, x} C(\tilde{x}, x)\|_2 - 1)^2]
 \end{aligned} \tag{2.6}$$

where $P_{(\tilde{x}, x)}$ is the joint probability of $\tilde{x} = \epsilon x + (1 - \epsilon)G(z, x_c)$ and x ; ϵ is sampled from a uniform distribution in $[0, 1]$; λ is the hyper-parameter that controls the gradient penalty.

Similar setup as SEGAN, the generator receives the noisy speech signal and latent vectors and put them into multi-layers convolutions with the filter (width = 31 and strides $N = 2$). Before the intermediate layer, a normal 2D convolutional followed by Parametric Rectified Linear Units (PReLU) [22] is used for inherent information capturing from input distributions. Then 2D transposed convolutional, followed again by PReLU, is used for information reconstruction.

The discriminator is considered as a binary classifier for judgement to real samples and generated samples. The main component is the 2D convolutional layer as well.

Differently, the discriminator applies the LeakyReLU function [13] and virtual batch normalization function rather than only PReLU in the generator. This will greatly improve the discriminable information learning of the discriminator and alleviate gradients vanishing.

The main structural ingredients of the encoder are also the 2D convolutional layer. Especially, the multi-head self-attention layer is applied to the encoder for the specific speech information locating and the long-range dependencies learning [32].

2.3.2 Multi-head Self-attention

For each input sequence, the Query (Q), Key (K), and Value (V) vectors will be created by applying learned linear projection or using feed-forward layers. Then the attention will be applied to all other positions with the three vectors. The procedure can be described as below:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \quad (2.7)$$

where d_k is the dimension of the key vectors. The purpose of this scaling is to improve numerical stability as the dimensions of keys, values, and queries grow. The obtained attention at each position will be used to times the value vector of all other positions including itself. This will produce multiple results called multi-head attention. The sum of all heads will be the final result of the first position input. The same operation will be applied at each subsequent position. Below is the equation of the multi-head calculation:

$$MultiHead(Q, K, V) = \text{concat}(head_i, \dots, head_h)W^O \quad (2.8)$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, the matrices W_i^Q , W_i^K , W_i^V , and W^O are the projection weight matrices, respectively. The self-attention module can calculate the response at a specific local position based on the resource collecting from all positions, where the attention vectors are calculated with a small computational cost.

2.4 Experiments

2.4.1 Database

The selected database is an open and standard resource for the performance evaluation of a speech enhancement system. The original clean speech was selected from Voice Bank corpus¹, including 28 speakers – 14 males and 14 females with the same accent region (England). There are two artificially generated noises (i.e. speech-shaped noise and babble) and eight real noises extracted from the Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) database [30].

For training data, the Signal-to-Noise Ratio (SNR) values are 15dB, 10dB, 5dB and 0dB. That signifies 40 different noisy conditions are produced in this corpus. Each speaker contributes 10 sentences, the corpus will add 400 sentences in total. Each clean speech waveform needs to be normalized and trimmed off silence segments of which are longer than 200ms at the beginning and the end.

¹<https://datashare.is.ed.ac.uk/>

Another two speakers (a male and a female), not including in the training data, are picked as the test data from the Voice Bank corpus with the same England accent. Five other noisy types were selected from the DEMAND database. The SNR values are 17.5dB, 12.5dB, 7.5dB and 2.5dB, respectively. Thereby, there are 20 different conditions for each sentence per test speaker.

2.4.2 Setup

ALRL adopts the Adam optimizer [3], a learning rate of 0.0002. The raw speech waveforms preserving the original inherent content of speech signals are used the same as SEGAN [25]. About one-second speech chunks (16384 samples) is segmented by a sliding window (500ms overlap) during training, however, is no overlap during the test. In addition, a high-frequency pre-emphasis filter of coefficient 0.95 to all input samples is applied. The epoch is 80 and the batch size is 100. In this work, fully convolution is used for distribution modelling during downsampling. For more stable training, the 2D convolutional layers followed by PReLU [22] are applied to the project and compress the input signal. Furthermore, the 2D transposed convolutional layers are designed as the key components of upsampling to reconstruct condensed representations.

2.5 Results

Many objective evaluation measures can evaluate enhanced speech performance with high correlation. The Perceptual Evaluation of Speech Quality (PESQ: from -0.5 to 4.5) for wide band speech is an effective full-reference speech quality evaluation algorithm [15], Moreover, we also implement the composite evaluation metrics of the enhanced speech including the predicted Mean Opinion Score (MOS) of signal

distortion (CSIG: from 1 to 5), background noise distortion (CBAK: from 1 to 5), and overall quality (COVL: from 1 to 5).

The intelligibility of enhanced speech is also implemented in this work. The Coherence-based Speech Intelligibility Index (CSII) measure is computed for the medium-level (CSII_{mid}) and high-level (CSII_{high}) segments of each speech sentence, which can predict the intelligibility of peak-clipping and centering-clipping distortion in the speech signal [21]. In addition, another popular speech intelligibility evaluation metrics the Normalized Covariance Metric (NCM) [21] and the Short-Time Objective Intelligibility (STOI) [28] are also conducted.

The experimental results of different methods are shown in Table 2.1. We set the SEGAN method as the baseline and its result was described in study [25]. According to the description of the SERGAN method [2], we retrain the SERGAN model and obtained the results as shown in Table 2.1. Besides the overall evaluation results, we also split the test data to four respective SNR conditions (i.e. 17.5dB, 12.5dB, 7.5dB, and 2.5dB) and obtain evaluation results.

As shown in Table 2.1, our Adversarial Latent Representation Learning (ALRL) method outperforms the SEGAN and SERGAN methods and achieves the highest scores in both speech quality and intelligibility. Specifically, our method improves PESQ by 1.98% and 19.0%, improves STOI by 0.213‰ and 9.81‰ over SERGAN and SEGAN, respectively. Moreover, our method also obtains outstanding enhancement performance in each SNR condition. Our method improves PESQ by 1.69% and 15.4% in 17.5dB, 2.39% and 21.6% in 2.5dB, improves STOI by -0.208‰ and 7.23‰ in 17.5dB, 1.33‰ and 14.4‰ in 2.5dB over SERGAN and SEGAN. Our ALRL can effectively improve the intelligibility and quality of noisy speech, especially for low SNR scenarios.

TABLE 2.1: The evaluation results of different methods in quality and intelligibility. The results include four SNR conditions (i.e. 17.5 dB, 12.5dB, 7.5dB, and 2.5 dB) and the overall values. "†" denotes that we reproduced the results with the provided open resources. The best scores are highlighted in bold.

Strategies		Quality					Intelligibility(%)			
Methods	SNR	PESQ	CSIG	CBAK	CVOL	CSII _{high}	CSII _{mid}	NCM	STOI	
SEGAN†	17.5dB	2.60	3.93	3.28	3.26	99.67	95.60	99.39	95.43	
	12.5dB	2.29	3.65	3.06	2.96	99.12	91.10	98.91	94.25	
	7.5dB	2.06	3.36	2.87	2.69	97.68	85.21	97.18	92.91	
	2.5dB	1.76	3.02	2.59	2.35	93.10	74.78	92.81	89.03	
	Overall	2.16	3.48	2.94	2.80	97.29	86.37	96.98	92.80	
SERGAN†	17.5dB	2.95	4.10	3.56	3.51	99.75	96.93	99.69	96.14	
	12.5dB	2.67	3.81	3.31	3.21	99.36	93.36	99.41	95.70	
	7.5dB	2.43	3.57	3.09	2.97	98.21	87.89	98.49	93.73	
	2.5dB	2.09	3.22	2.79	2.61	94.63	78.38	95.91	90.19	
	Overall	2.52	3.66	3.18	3.06	97.91	88.89	98.33	93.69	
ALRL	17.5dB	3.00	4.21	3.65	3.60	99.76	96.98	99.71	96.12	
	12.5dB	2.73	3.94	3.36	3.32	99.39	93.50	99.40	95.11	
	7.5dB	2.47	3.69	3.13	3.06	98.28	88.03	98.61	93.67	
	2.5dB	2.14	3.32	2.82	2.70	94.81	78.80	96.20	90.31	
	Overall	2.57	3.78	3.23	3.16	97.98	89.08	98.43	93.71	

2.6 Conclusions

In this chapter, we propose a novel Adversarial Latent Representation Learning (ALRL) method for speech enhancement. An extra encoder model is built in our ALRL to learn the semantic representation by inverse mapping from the generated samples to the latent space. The encoder greatly improves the effectiveness of adversarial training and complex data distribution learning. To accomplish the inverse mapping, we propose a new loss function, which captures latent representation by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors. In addition, the multi-head self-attention mechanism, applied to the encoder, is also effective for long-range dependencies capturing and further semantic representation learning. The experimental results demonstrate that ALRL outperforms current existing methods in both speech quality and intelligibility, especially for low signal-to-noise ratio scenarios. Our experiments have shown that the latent space is effective to learn semantic representation with adversarial training and our ALRL is effective for speech enhancement performance improvement.

References

- [1] Deepak Baby. iSEGAN: Improved speech enhancement generative adversarial networks. *arXiv preprint arXiv:2002.08796*, 2020.
- [2] Deepak Baby and Sarah Verhulst. SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 106–110. IEEE, 2019. doi: <https://doi.org/10.1109/ICASSP.2019.8683799>.

-
- [3] Irwan Bello, Barret Zoph, Vijay Vasudevan, and Quoc V Le. Neural optimizer search with reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 459–468, 2017.
 - [4] Danyang Cao, Zhixin Chen, and Xue Gao. Research on noise reduction algorithm based on combination of lms filter and spectral subtraction. *Journal of Information Processing Systems*, 15(4), 2019. doi: <https://doi.org/10.3745/JIPS.04.0123>.
 - [5] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2172–2180, 2016.
 - [6] Fu-Kai Chuang, Syu-Siang Wang, Jeh-wei Hung, Yu Tsao, and Shih-Hau Fang. Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3173–3177, 2019. doi: <https://doi.org/10.21437/Interspeech.2019-2108>.
 - [7] Chris Donahue, Bo Li, and Rohit Prabhavalkar. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5024–5028. IEEE, 2018. doi: <https://doi.org/10.1109/ICASSP.2018.8462581>.
 - [8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.
 - [9] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 658–666, 2016.
 - [10] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex

- Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.
- [11] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee. Densely connected progressive learning for LSTM-based speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5054–5058. IEEE, 2018. doi: <http://doi.org/10.1109/ICASSP.2018.8461861>.
- [12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. doi: <http://doi.org/10.1109/ICCV.2015.123>.
- [14] Tsun-An Hsieh, Hsin-Min Wang, Xugang Lu, and Yu Tsao. Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement. *IEEE Signal Processing Letters*, 27:2149–2153, 2020. doi: <http://doi.org/10.1109/LSP.2020.3040693>.
- [15] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(1):229–238, 2007. doi: <http://doi.org/10.1109/TASL.2007.911054>.
- [16] Alexia Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard GAN. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.
- [17] Mathew Shaji Kavalekalam, Jesper Kjar Nielsen, Jesper Bunsow Boldt, and

- Mads Græsboll Christensen. Model-based speech enhancement for intelligibility improvement in binaural hearing aids. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(1):99–113, 2019. doi: <https://doi.org/10.1109/TASLP.2018.2872128>.
- [18] Ju Lin, Sufeng Niu, Zice Wei, Xiang Lan, Adriaan J van Wijngaarden, Melissa C Smith, and Kuang-Ching Wang. Speech enhancement using forked generative adversarial networks with spectral subtraction. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3163–3167, 2019. doi: <https://doi.org/10.21437/Interspeech.2019-2954>.
- [19] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. 2013. ISBN 1466504218. doi: <https://doi.org/10.1201/9781420015836>.
- [20] Siow Yong Low. Compressive speech enhancement in the modulation domain. *Speech Communication*, 102:87–99, 2018. doi: <https://doi.org/10.1016/j.specom.2018.08.003>.
- [21] Jianfen Ma, Yi Hu, and Philipos C Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387–3405, 2009. doi: <https://doi.org/10.1121/1.3097493>.
- [22] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.
- [23] Daniel Michelsanti and Zheng-Hua Tan. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2008–2012, 2017. doi: <https://doi.org/10.21437/Interspeech.2017-1620>.
- [24] Zhiheng Ouyang, Hongjiang Yu, Wei-Ping Zhu, and Benoit Champagne. A fully

- convolutional neural network for complex spectrogram processing in speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5756–5760. IEEE, 2019. doi: <http://doi.org/10.1109/ICASSP.2019.8683423>.
- [25] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: Speech enhancement generative adversarial network. In *Proceedings of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3642–3646, 2017. doi: <http://doi.org/10.21437/Interspeech.2017-1428>.
- [26] Pourya Shamsolmoali, Masoumeh Zareapoor, Ruili Wang, Deepak Kumar Jain, and Jie Yang. G-GANISR: Gradual generative adversarial network for image super resolution. *Neurocomputing*, 366:140–153, 2019. doi: <https://doi.org/10.1016/j.neucom.2019.07.094>.
- [27] Pourya Shamsolmoali, Masoumeh Zareapoor, Eric Granger, Huiyu Zhou, Ruili Wang, M Emre Celebi, and Jie Yang. Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion*, 2021. doi: <https://doi.org/10.1016/j.inffus.2021.02.014>.
- [28] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 19(7):2125–2136, 2011. doi: <https://doi.org/10.1109/TASL.2011.2114881>.
- [29] Naohiro Tawara, Tetsunori Kobayashi, and Tetsuji Ogawa. Multi-channel speech enhancement using time-domain convolutional denoising autoencoder. In *Proceedings of 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 86–90, 2019. doi: <https://doi.org/10.21437/Interspeech.2019-3197>.
- [30] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database: A database of multichannel

- environmental noise recordings. *The Journal of the Acoustical Society of America*, 133(5):3591–3591, 2013. doi: <https://doi.org/10.1121/1.4806631>.
- [31] Yan-Hui Tu, Jun Du, and Chin-Hui Lee. Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(12):2080–2091, 2019. doi: <https://doi.org/10.1109/TASLP.2019.2940662>.
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.
- [33] Yan Yang and Changchun Bao. Dnn-based AR-Wiener filtering for speech enhancement. In *Proceedings of the 43th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2901–2905. IEEE, 2018. doi: <https://doi.org/10.1109/ICASSP.2018.8462563>.

Chapter 3

Adversarial Multi-Task Learning with Inverse Mappings for Speech Enhancement

For speech enhancement, Adversarial Multi-Task Learning (AMTL) has demonstrated its promising capability of information capture and representation learning in complex scenarios. However, previous AMTL-based approaches focused on enhancing the distinguishable performance of the discriminator. In this chapter, we propose a novel Adversarial Multi-Task Learning with Inverse Mappings (AMTL-IM) method for speech enhancement. Our method focuses on enhancing the generator's capability of speech information capture and representation learning. To implement our method, two extra networks (namely P and Q) are developed to establish the inverse mappings from the generated distribution to the input data domains. Correspondingly, the latent loss and equilibrium loss are proposed for the inverse mappings learning and the enhancement model training based on the original adversarial loss.

3.1 Introduction

Speech enhancement, one of the most important topics in speech signal processing [33], is to improve the intelligibility and overall perceptual quality of degraded speech. The intelligibility is a measurement of how comprehensible a speech is, while the perceptual quality measures how easy it is for a listener to perceive the content of a speech. Normally, a perceptual high-quality speech sounds more natural, rhythmic, yet less raspy, hoarse, or scratchy. In practice, speech enhancement has been widely applied in scenarios such as mobile communications [3], hearing aids [27], and noise-robust speech recognition or speaker recognition [51, 54]. In real life, there are various negative interferences such as additive noise (e.g., fan noise) and convolutional noise (e.g., room reverberation), which can badly degrade speech intelligibility and overall perceptual quality.

Different speech enhancement methods have been proposed to eliminate the negative effects of the environmental noises [1, 9, 29, 38, 43, 45, 46, 59, 60]. For example, Wiener filtering is a classic single-channel statistical estimation based approach, which is considered to be an effective way for stationary additive noise reduction [1]. However, for reverberation or unknown noise interference, the statistical estimation based approaches perform unsatisfactorily in the current complex noisy environment.

Another popular approach is microphone arrays based multi-channel speech enhancement [29, 43], such as the acoustical beamforming algorithm [61]. This approach is conducted on the output signals of microphone arrays and converts them into a single-channel speech signal while amplifying the speech signals from the targeted direction and attenuating the noise signals coming from other directions. The microphone arrays based multi-channel approaches usually take the spatial position information into account and can effectively mitigate the reverberation problem [29].

With the rapid development of intelligent technologies and hardware resources, data-driven deep neural networks have been thriving in speech signal processing [52], computer vision [6, 25, 47], and natural language processing [20]. For speech enhancement, denoising autoencoder [7, 53], Long Short-Term Memory (LSTM) [16], and Convolutional Neural Networks (CNN) based methods [21, 40] have been applied to improve high-dimensional data representation learning and speech enhancement performance as well. For example, Zhao et al. [62] introduced convolutional-recurrent neural networks to exploit local structures in the frequency and temporal domains. The results showed that their method was more data-efficient and achieved better generalization on both seen and unseen noise. With a deeper neural architecture, deep learning based approaches display a huge potential in dealing with complex signal processing and specific representation learning [39]. However, recent works also revealed that the performance of deep architecture degrades inversely if we exhaustively enlarge the network’s scale only [28], which would cause vanishing gradients or degradation problems. Benefiting from the normalized initialization [18], intermediate normalization layers [24], and skip connections of residual network [19], these aforementioned problems can be largely addressed.

Recently, the advent of Generative Adversarial Networks (GAN) [17] has attracted much attention and made remarkable progress in the generative model community. With the powerful ability of information learning and reconstruction, complex image and speech signal processing achieved a great breakthrough [2, 47]. The original GAN consists of a Generator (G) and a Discriminator (D). G is set to learn an effective mapping between the given random noise (z) and the ground-truth (x). In contrast, D is an initialized binary classifier, which receives both x (real) and $G(z)$ (fake) and gives a corresponding judgment. With continuous iterative processing, the procedure is trained adversarially up to a Nash Equilibrium [17]. For speech enhancement [41] as shown in Figure 3.1, G usually takes noisy speech and extra

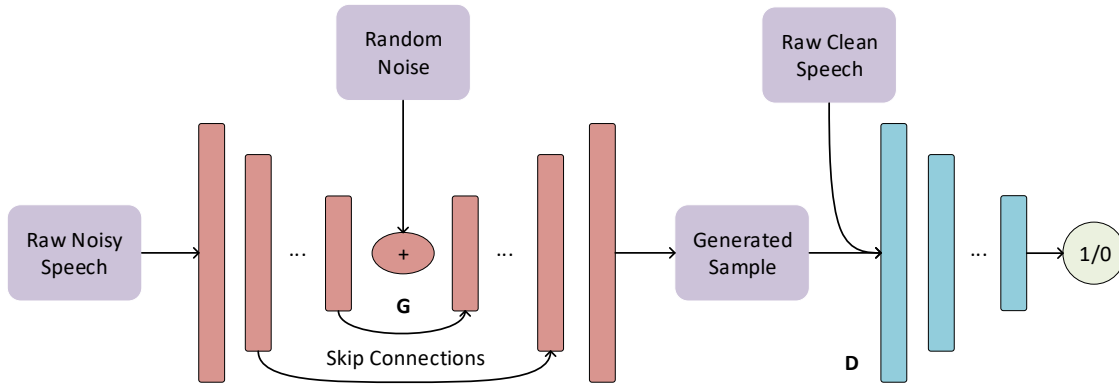


FIGURE 3.1: The framework of GAN-based speech enhancement. The Generator (G) consumes raw noisy speech and latent vector (i.e., random noise) as input. The Discriminator (D) is a binary classifier aiming to judge the similarity between the generated sample and raw clean speech.

noise distributions as input and exports targeted data distribution. D is considered as a classifier trained to distinguish generated samples and clean speech as fake or real.

Although numerous speech enhancement methods have been developed in the past and worked properly in many scenarios to some extent, it is necessary to further improve the generalization of the model and the performance of speech enhancement. Adversarial Multi-Task Learning (AMTL), which combines an adversarial training mechanism with Multi-Task Learning (MTL), is an effective method to improve the complex multiple domains information learning and generalization [15, 63]. Specifically, MTL is an inductive transfer mechanism aiming to improve the generalization performance by leveraging the domain-specific information contained in related tasks [4].

With powerful information capture and reconstruction capability, AMTL has been used in text classification [31], image feature learning [32], speaker normalization in replay detection [49] and speech enhancement [35, 36]. For speech enhancement, Meng et al. [35, 36] proposed AMTL-based methods to enhance the distinguishable performance of models. In [36], two discriminators were added on top of the basic

cycle-consistent framework. The multiple losses including the discrimination losses, the reconstruction losses, and the identity-mapping losses were jointly optimized to distinguish the enhanced and noised features from the real samples. A similar AMTL-based idea was proposed in [35]. The experimental results showed that Meng’s AMTL-based speech enhancement methods effectively reduced the Word Error Ratio (WER) of noise-robust speech recognition. The AMTL-based methods of Meng et al. [36] focused on discriminability improvement of the discriminator. However, little attention has been paid to improve the specific information capture and speech representation learning of generators.

Moreover, establishing the inverse mappings from the output distribution domain to the input data domain is an effective way to improve representation learning [11, 12, 23]. Thus, in this chapter, we propose a novel Adversarial Multi-Task Learning with Inverse Mappings (AMTL-IM) method for speech enhancement. Based on the architecture of GAN, two extra networks (namely P and Q) are developed to establish the inverse mappings from the generated distribution to the input data domains. Correspondingly, two new loss functions (i.e., latent loss and equilibrium loss) are proposed for the inverse mappings learning and the enhancement model training based on the original adversarial loss. With the latent loss function, network P aims to extract relevant latent information from the latent space (i.e., random noise domain) and further facilitate the sample generation. The network Q is developed to balance the adversarial representation learning by mapping the generated distribution to the noisy speech domain with an equilibrium loss function. Thus, our proposed method consists of four sub-models with respective loss functions to learn speech representation and further improve speech enhancement.

This chapter is organized as follows. Section 3.2 describes the related work. Section 3.3 details the proposed method. Section 3.4 provides the details of our experiments.

The results and discussions are presented in Section 3.5. Finally, the conclusions are shown in Section 3.6.

3.2 Related Work

In this section, we introduce the related work about GAN-based speech enhancement and inverse mapping learning.

3.2.1 Speech Enhancement with Adversarial Networks

Recently, GAN-based models have achieved huge progress on semantic representation learning and improved speech enhancement performance significantly. Speech Enhancement GAN (SEGAN) [41] is one of the most prominent frameworks proposed for time-domain speech enhancement. SEGAN combines the conditional GAN with the Least-Squares GAN (LSGAN) to further alleviate vanishing gradients. This modification is proved to be effective for performance improvement. Below is the loss function of its discriminator:

$$L_D = \frac{1}{2} E_{x \sim P_x, x_c \sim P_{x_c}} [(D(x, x_c) - 1)^2] + \frac{1}{2} E_{z \sim P_z, x_c \sim P_{x_c}} [(D(G(z, x_c), x_c))^2], \quad (3.1)$$

and its generator:

$$L_G = \frac{1}{2} E_{z \sim P_z, x_c \sim P_{x_c}} [(D(G(z, x_c), x_c) - 1)^2], \quad (3.2)$$

where x_c denotes noisy speech; x denotes clean speech; z denotes random noise distribution.

SEGAN operated directly on the raw speech waveform rather than on the processed spectral features with an end-to-end architecture. The end-to-end architecture is

considered to be able to preserve original sequential information such as phase information effectively. The fully convolutional architecture consisted of a downsampling module and an upsampling module (i.e., encoder and decoder). This enforced the network to focus on temporally close correlations of the input speech signal and throughout the entire processing of the network [41]. The random noise z (i.e., latent vectors) was added to the bottleneck layer for information compensation. However, the latent space has not been explored thoroughly in speech signal processing.

The latent vectors may be used by the generator in a highly entangled way [5]. For inducing latent vectors, Chen et al. [5] proposed to adopt a mutual information strategy, which decomposed the input noise vectors into a set of semantically meaning factors of variation rather than using single unstructured noise vectors. They discovered that these latent factors could target salient semantic features of data distribution. Thus, establishing an inverse mapping to explore the latent space for effective representation learning is one of the main tasks in our work.

Speech Enhancement Relativistic GAN (SERGAN) [2] is another framework exploring speech enhancement based on GAN. In the conventional GAN, the discriminator is trained to detect if a sample is an original one or a generated one, while the generator is trained to generate data to be more similar to original data to fool the discriminator. The relativistic GAN [26] argued that the probability of $D(x)$ should decrease as the probability of $D(G(z, x_c))$ increases. However, the original GAN cannot incorporate this situation described above since G does not influence $D(x)$. To circumvent this problem, the relativistic loss function was proposed and used in the speech enhancement task [2]. Below is the loss function of SERGAN’s discriminator

:

$$L_D = -E_{x \sim P_x, x_c \sim P_{x_c}}[\log(\sigma(C(x, x_c) - C(G(z, x_c), x_c)))], \quad (3.3)$$

and generator:

$$L_G = -E_{x \sim P_x, x_c \sim P_{x_c}} [\log(\sigma(C(G(z, x_c), x_c) - C(x, x_c)))], \quad (3.4)$$

where $C(x)$ denotes the discriminator without the final sigmoid layer; σ is the sigmoid non-linearity, and thus $D(x) = \sigma(C(x))$. The method had a similar architecture with SEGAN but adopted a new loss function to boost information communication between the generator and discriminator.

Benefiting from adversarial training, more GAN-based methods have been proposed for speech enhancement. Michelsanti and Tan [37] explored the potential of the conditional GAN for speech enhancement; Soni et al. [48] exploited GAN with time-frequency mask based enhancement framework; Donahue et al. [10] conducted a detailed study to measure the effectiveness of GAN-based speech enhancement for robust speech recognition where the speech is contaminated by both additive and convolutional noise.

With various model architectures and task requirements, GAN-based speech enhancement has demonstrated a promising capability of complex distribution modelling and speech representation learning.

3.2.2 Inverse Mapping Learning

For effective information capture and data representation learning, the inverse mappings learning with GAN’s architecture has shown its success in image processing [11, 12, 23]. The Stacked GAN (SGAN) [23] was proposed to invert the hierarchical representations of a traditional bottom-up encoder to a stack of top-down generators for high-quality image generation. Each generator learned to generate lower-level data representations conditional upon high-level representations. The bottom-up

encoder, employing a fully connected network, was pre-trained to provide inherent information for the stacked layers of generators. The separated generators and discriminators were trained independently and then jointly to invert the hidden layers information of the encoders. The iterative adversarial training transformed the inherent information from the bottom-up encoder to the top-down generators for higher resolution image output. The proposed method improved the inherent information learning and enhanced high-resolution image generation by inverting the hidden layers information to the target data domain inversely.

Donahue et al. [11] noticed that GAN models could capture semantic variation from latent space but with no means of projecting data back into the latent space. Thus, the GAN architecture ignored much of the useful information found in the structure of the data itself. Besides, interpolations in the latent space of the generator produced smooth and plausible semantic variations and made the model learn to associate particular latent directions with specific features. Thus, the Bidirectional Generative Adversarial Networks (BiGAN) was proposed to learn an inverse mapping from the projecting data back into the latent spaces [11] with a new encoder model. The learned feature representation was thus useful for auxiliary supervised discrimination tasks. Another similar work about latent space exploration with multiple models was proposed in [13]. As introduced in [13], the generation network mapped samples from stochastic latent variables to the data domain while the inference network mapped training examples in the data domain to the space of latent variables inversely. The operation could effectively enhance representation learning and sample reconstruction with an adversarial process.

The related work mentioned above demonstrates that the GAN-based methods have a huge potential in speech enhancement tasks. To further improve the performance of GAN-based speech enhancement, exploring the input data domain with inverse

mappings is an effective way. Thus, in this chapter, we propose a novel speech enhancement method based on adversarial multi-task learning and inverse mappings.

3.3 Methodology

Based on adversarial multi-task learning and inverse mapping learning, we propose a novel method to further enhance speech representation learning and the performance of speech enhancement. As shown in Figure 3.2, our method consists of four networks: a generator G , a discriminator D , and the proposed two extra networks P and Q . G consumes raw noisy speech and random noise as input and outputs the generated sample; D aims to judge the similarity between the generated sample and the raw clean speech. The networks P and Q compensate the information extracted by G by learning the inverse mappings from the generated distribution to the input spaces with two additional new loss functions (i.e., latent loss and equilibrium loss).

3.3.1 Loss Functions

The loss function of G consists of three parts: adversarial loss ($L_{G.adv}$), latent loss ($L_{G.lat}$), and equilibrium loss ($L_{G.equ}$). The weighted sum of these three parts is expected to capture real-data information and learn an effective representation of G . Below is the combined loss function of G :

$$L_G = L_{G.adv} + \lambda_1 L_{G.lat} + \lambda_2 L_{G.equ}, \quad (3.5)$$

where P and Q can be activated or deactivated by setting λ_1 and λ_2 as 0 or 1. We can also try different weight groups to evaluate the effect of P and Q on the whole model.

The basic adversarial loss function can learn necessary information for G when D is frozen. Here, we adopt the adversarial loss function used in SERGAN [2]. Below is the adversarial loss function:

$$L_{G_{adv}} = -E_{x \sim P_x, x_c \sim P_{x_c}} [\log(\sigma(C(G(z, x_c), x_c) - C(x, x_c)))], \quad (3.6)$$

where x denotes the clean speech, which subjects to data distribution P_x ; x_c denotes the noisy speech, which subjects to data distribution P_{x_c} ; z denotes the random noise subjecting to distribution P_z (i.e., latent space); $D(x, x_c) = \sigma(C(x, x_c))$ as mentioned above.

The latent space plays an important role in GAN architecture-based representation learning and stable model training [11]. However, current models generally ignore thoroughly exploring latent space information. Thus, in our method, P is built to excavate latent space information by mapping the generated distribution to the latent space inversely. Below is the latent loss function:

$$L_{G_{lat}} = -E_{z \sim P_z, x_c \sim P_{x_c}} [\|P(G(z, x_c)) - z\|_2^2], \quad (3.7)$$

where the squared Euclidean distance $\|\cdot\|_2^2$ is adopted to measure the similarity of random noise distribution z with the output distribution of P . Here, the distance measurement can be designed in other ways, but we choose $\|\cdot\|_2^2$ because it makes the hyper-parameter tuning easier [12].

We propose to establish the inverse mapping from generated data distribution to latent space with network P . The latent loss function works with the adversarial loss function together to enhance G to capture more effective information for information reconstruction. However, a potential unbalanced learning problem may appear

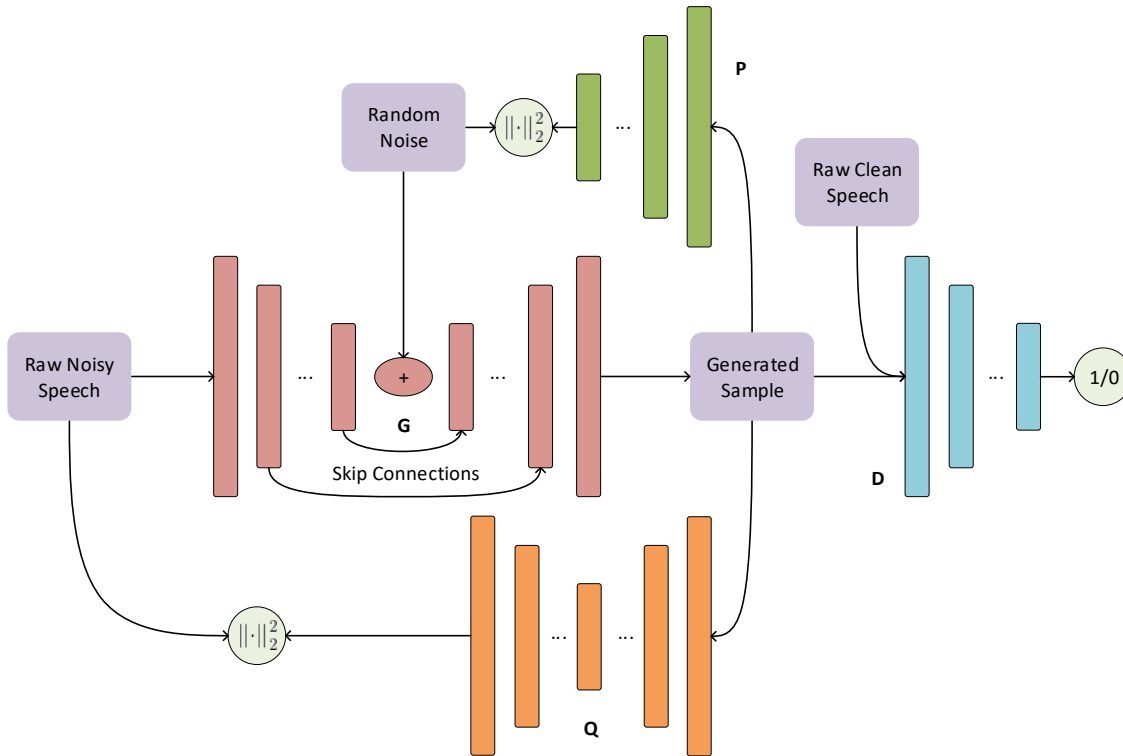


FIGURE 3.2: The framework of our proposed method. Based on basic GAN architecture, the Generator (G) receives raw noisy speech and random noise as input. The Discriminator (D) gives the judgment (i.e., fake or real) of the generated sample and raw clean speech. The networks P and Q are proposed to establish the inverse mappings from the generated distribution to the input data domain for information capture and representation learning.

and result in defective real-data distribution modelling. Also, unnecessary complexity and model instability may be introduced if just feeding more extra conditional information [23]. Thus, another network Q with the equilibrium loss function is developed as well to obtain a trade-off during model training. Below is the equilibrium loss function:

$$L_{G_equ} = -E_{z \sim P_z, x_c \sim P_{x_c}} [\|Q(G(z, x_c)) - x_c\|_2^2], \quad (3.8)$$

where $\|\cdot\|_2^2$ is also adopted to measure the similarity of noisy speech distribution with the output of Q .

With the pre-set weights, the adversarial loss, latent loss, and equilibrium loss functions form the overall loss function of G . The multi-task learning based architecture achieves the related distribution mapping and representation learning with an adversarial training mechanism.

Following [2], the gradient penalty regularization is also applied in our work to avoid further vanishing gradients. Below is the loss function of D :

$$L_D = - E_{x \sim P_x, x_c \sim P_{x_c}} [\log(\sigma(C(x, x_c) - C(G(z, x_c), x_c)))] \\ - \gamma E_{\tilde{x}, x \sim P_{(\tilde{x}, x)}} [(\|\nabla_{\tilde{x}, x} C(\tilde{x}, x)\|_2 - 1)^2] \quad (3.9)$$

where x and x_c denote the clean and noisy speech pair, which subject to data distribution P_x and P_{x_c} , respectively; z denotes random noise, which subjects to P_z ; $P_{(\tilde{x}, x)}$ is the joint probability of $\tilde{x} = \epsilon x + (1 - \epsilon)G(z, x_c)$ and x ; ϵ is sampled from a uniform distribution in $[0, 1]$, and $\gamma = 10$ is the hyper-parameter that controls the gradient penalty.

3.3.2 Model Architecture Setup

In this subsection, we introduce the architecture of our model. As shown in Figure 3.3, G is a standard downsampling and upsampling architecture developed for information learning and reconstruction. Before the intermediate bottleneck layer, the normal 2D convolutional kernels followed by Parametric Rectified Linear Units (PReLU) [18] are adopted for information capture from real-data distributions. Then, the 2D transposed convolutional kernels with PReLU are applied for desirable sample reconstruction. Latent vector z gets concatenated with the condensed representation of the bottleneck layer. Additionally, the skip connections linking the

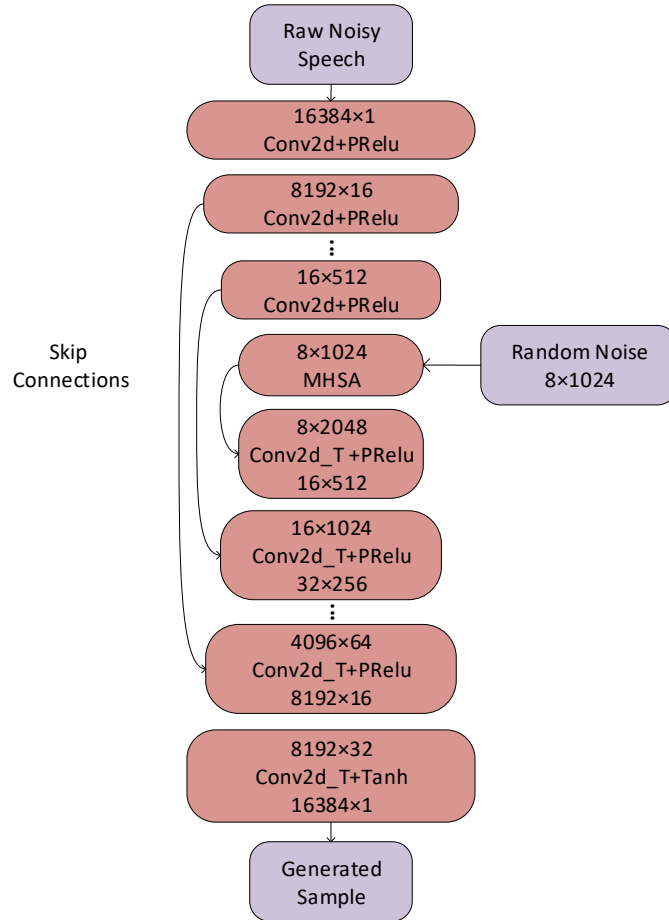


FIGURE 3.3: The details of Generator (G). The downsampling adopts 2D convolutional kernels followed by PReLU for information capture. The upsampling adopts 2D transposed convolutional kernels followed by PReLU for sample reconstruction. Latent vector z gets concatenated with the condensed representation of the bottleneck layer. The skip connections are used to boost the stability of model training.

downsampling and upsampling of G can transfer the fine-grained information of the speech waveform to the upsampling stage and boost the stability of model training [42].

As shown in Figure 3.4, D is considered as a binary classifier to judge the similarity between the ground truth and the generated sample. The main components of D are also the 2D convolutional kernels but followed by Virtual Batch Normalization

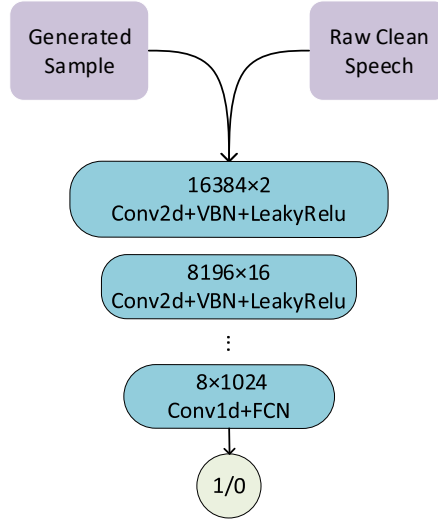


FIGURE 3.4: The details of Discriminator (D). D is a binary classifier to give the judgment (fake or real) of the ground truth and the generated sample. The main components are the 2D convolutional kernels followed by Virtual Batch Normalization (VBN) and LeakyReLU for distinguishable information learning.

(VBN) and LeakyReLU [41]. This architecture is suitable for learning distinguishable information.

In this work, network P also employs 2D convolutional kernels with PReLU similar to D but removes the final fully connected layer to match the dimension of random noise. Network Q employs a downsampling and upsampling architecture similar to G but reduces the number of layers and discards the skip connections. The model’s training procedure is presented in Algorithm 1. In particular, we also apply the multi-head self-attention to G and Q in the bottleneck layer for locating specific speech information and learning the contextual long-range dependencies.

Algorithm 1 Training procedure of our speech enhancement method

Require:

Raw clean-noisy speech pairs (x, x_c) and random noise z
 Initialized weights and biases of D, G, P, Q networks

Ensure:

Trained speech enhancement model
 1: $\theta_D, \theta_G, \theta_P, \theta_Q \leftarrow$ initialize network parameters
 2: **for** epoch (number of training iterations) **do**
 3: speech signal pre-processing
 4: $(z, x_c) \leftarrow$ batch input of G
 5: $G(z, x_c) \leftarrow$ enhanced output of G
 6: $(x, G(z, x_c)) \leftarrow$ batch input of D
 7: $L_D, L_G \leftarrow$ loss calculation
 8: $\theta_D, \theta_G, \theta_P, \theta_Q \leftarrow$ parameters update
 9: **end for**
 10: **return** trained model

3.4 Experiments

3.4.1 Database

The selected database [55, 56] is an open and standard speech corpus for the evaluation of speech enhancement systems. The database contains selected speech resources from multiple speech corpus. Some of the noise files were obtained from the DEMAND corpus¹. Another two noise files² (i.e., the speech-shaped and babble noise) were also selected for noisy speech production. The original clean speech was selected from the Voice Bank corpus [58]. According to the number of speakers, two sub-databases were created: one includes 28 speakers (14 males and 14 females) with the same accent (England); another one includes 56 speakers (28 males and 28 females) with different accents (Scotland and United States).

As mentioned above, the database added ten different noise types to the clean speech waveform using the ITU-T P.56 method [33], including eight real noise types and two

¹<http://parole.loria.fr/DEMAND/>

²<http://homepages.inf.ed.ac.uk/cvbotinh/se/noises/>

TABLE 3.1: The evaluation results of our method compared with previous methods including Wiener filtering [30], SEGAN [41], SERGAN [2], MMSE-GAN [48], BiLSTM [14], CRN-MSN [52], NAAGN [8]. All the presented methods were trained with the 28-speaker database. "†" denotes that we reproduced the results with the provided open resource. "-" denotes that the result is not reported or not available. The best scores are highlighted in bold.

Models	PESQ	CSIG	CBAK	CVOL	SSNR	STOI
Noisy	1.97	3.35	2.44	2.63	1.68	0.921
Wiener filtering	2.22	3.23	2.68	2.67	5.07	-
SEGAN†	2.16	3.48	2.94	2.80	7.73	0.928
SERGAN†	2.52	3.66	3.18	3.06	9.40	0.937
MMSE-GAN	2.53	3.80	3.12	3.14	-	0.930
BiLSTM	2.70	3.99	2.95	3.34	-	0.925
MDPhD	2.70	3.85	3.39	3.27	10.2	-
CRN-MSE	2.74	3.86	3.14	3.30	-	0.934
NAAGN	2.90	4.13	3.50	3.51	10.3	0.948
Proposed $I_{(\lambda_1=1, \lambda_2=0)}$	2.57	3.78	3.23	3.16	9.32	0.937
Proposed $II_{(\lambda_1=0, \lambda_2=1)}$	2.52	3.73	3.22	3.11	9.06	0.935
Proposed $III_{(\lambda_1=1, \lambda_2=1)}$	2.79	3.90	3.34	3.56	9.67	0.941
Proposed $IV_{(\lambda_1=1, \lambda_2=1, MHS A)}$	2.88	4.01	3.50	3.51	9.72	0.945

artificially generated noises. In detail, the eight real noise types include a domestic kitchen room noise, a meeting room noise, three public space noises including cafeteria, restaurant, and subway station, two transportation noises including car and metro, a busy traffic intersection noise; the two artificially generated noises contain a speech-shaped noise by adding white noise, and a babble noise by adding extra speech.

For training data, the Signal-to-Noise Ratio (SNR) values were set to 15dB, 10dB, 5dB, and 0dB. It signified that each clean sentence would produce 40 noisy sentences with different noise types. Each speaker contributed with 10 clean sentences. Thus, each speaker would contribute with 400 sentences to the database. Moreover, each clean speech waveform would be normalized, and the silence segments would be trimmed off when the silence segments were longer than 200ms at the beginning and at the end.

Another two speakers (a male and a female), not included in the training data, were selected for the testing data with an accent from England. Five other noise types were selected from the DEMAND database, including a domestic living room noise, an office room noise, a transport noise of a bus, and two street noises including an open area and a public square. The SNR values were 2.5dB, 7.5dB, 12.5dB, and 17.5dB, respectively.

3.4.2 Setup

Our model is trained using the RMSprop optimizer [64] with a learning rate of 0.0002. The number of epochs is 180 and the batch size is 100. As an end-to-end architecture, our model takes in the raw speech waveform and outputs the enhanced waveform directly, which is considered to preserve the original content of speech signals including phase information. About one-second speech chunks (16384 samples) are segmented by a sliding window (500ms overlap) during training, however, there is no overlap during the test. Besides, a high-frequency pre-emphasis filter of coefficient 0.95 is applied to all input samples.

TABLE 3.2: The unfolded evaluation results on different SNR values (i.e., 17.5 dB, 12.5dB, 7.5dB, 2.5dB, and overall). We evaluate SEGAN [41], SERGAN [2], ALRL [44] and our method with more comprehensive speech quality and intelligibility metrics on the 28-speaker database. ”†” denotes that we reproduced the results with the provided open resources.

Strategies		Quality						Intelligibility			
Methods	SNR	PESQ	CSIG	CBAK	CVOL	SSNR	CSII _{high}	CSII _{mid}	CSII _{low}	NCM	STOI
SEGAN†	17.5dB	2.60	3.93	3.28	3.26	9.23	0.997	0.956	0.684	0.994	0.954
	12.5dB	2.29	3.65	3.06	2.96	8.46	0.991	0.911	0.587	0.989	0.943
	7.5dB	2.06	3.36	2.87	2.69	7.52	0.977	0.852	0.486	0.972	0.929
	2.5dB	1.76	3.02	2.59	2.35	5.88	0.931	0.748	0.338	0.928	0.890
	Overall	2.16	3.48	2.94	2.80	7.73	0.973	0.864	0.518	0.970	0.928
SERGAN†	17.5dB	2.95	4.10	3.56	3.51	11.7	0.998	0.969	0.738	0.997	0.961
	12.5dB	2.67	3.81	3.31	3.21	10.2	0.994	0.934	0.644	0.994	0.957
	7.5dB	2.43	3.57	3.09	2.97	8.83	0.982	0.879	0.561	0.985	0.937
	2.5dB	2.09	3.22	2.79	2.61	7.13	0.946	0.784	0.425	0.959	0.902
	Overall	2.52	3.66	3.18	3.06	9.40	0.979	0.889	0.587	0.983	0.937
ALRL	17.5dB	3.00	4.21	3.65	3.60	11.4	0.998	0.970	0.727	0.997	0.961
	12.5dB	2.73	3.94	3.36	3.32	10.1	0.994	0.935	0.637	0.994	0.951
	7.5dB	2.47	3.69	3.13	3.06	8.75	0.983	0.880	0.551	0.986	0.937
	2.5dB	2.14	3.32	2.82	2.70	7.16	0.948	0.788	0.423	0.962	0.903
	Overall	2.57	3.78	3.23	3.16	9.32	0.980	0.891	0.580	0.984	0.937
Proposed_IV	17.5dB	3.38	4.40	3.65	3.90	12.0	0.999	0.970	0.731	0.997	0.965
	12.5dB	3.15	4.20	3.36	3.66	10.5	0.996	0.938	0.651	0.996	0.957
	7.5dB	2.94	3.97	3.13	3.46	9.21	0.987	0.888	0.573	0.989	0.948
	2.5dB	2.48	3.51	2.82	3.10	7.40	0.957	0.795	0.443	0.967	0.914
	Overall	2.88	4.01	3.50	3.51	9.72	0.984	0.895	0.595	0.987	0.945

TABLE 3.3: The evaluation results of SEGAN [41], SERGAN [2] and our method with different scales of training database in terms of speech quality and intelligibility. “†” denotes that we reproduced the results with the provided open resources.

Methods	Strategies		Quality					Intelligibility				
	training data		PESQ	CSIG	CBAK	CVOL	SSNR	CSII _{high}	CSII _{mid}	CSII _{low}	NCM	STOI
SEGAN†	28spks		2.16	3.48	2.94	2.80	7.73	0.973	0.864	0.518	0.970	0.928
	56spks		2.46	3.63	3.10	3.03	7.83	0.977	0.879	0.554	0.973	0.934
	28+56spks		2.51	3.46	3.15	2.95	8.99	0.981	0.890	0.554	0.976	0.937
SERGAN†	28spks		2.52	3.66	3.18	3.06	9.40	0.979	0.889	0.587	0.983	0.937
	56spks		2.61	3.89	3.25	3.24	9.03	0.980	0.890	0.587	0.984	0.938
	28+56spks		2.60	3.79	3.28	3.18	9.64	0.981	0.893	0.593	0.983	0.938
Proposed _{IV}	28spks		2.88	4.01	3.50	3.51	9.72	0.984	0.895	0.595	0.987	0.945
	56spks		2.90	4.03	3.55	3.52	9.70	0.983	0.898	0.593	0.987	0.947
	28+56spks		2.93	4.08	3.48	3.55	9.46	0.983	0.895	0.596	0.988	0.946

3.4.3 Evaluation Metrics

Although subjective evaluation is more accurate and reliable, it is costly and time-consuming [22]. Many objective evaluation measures can evaluate enhanced speech with high correlation. The Perceptual Evaluation of Speech Quality (PESQ: from -0.5 to 4.5) for wideband speech is an effective full-reference speech quality evaluation algorithm [22]. Moreover, we also implement the composite evaluation metrics of the enhanced speech including the predicted Mean Opinion Score (MOS) of signal distortion (CSIG: from 1 to 5), background noise distortion (CBAK: from 1 to 5), and overall quality (COVL: from 1 to 5). The Segmental Signal-to-Noise Ratio (SSNR: from 0 to ∞) is another crucial evaluation metric for speech quality.

The intelligibility of enhanced speech is also tested in this work. The Coherence-based Speech Intelligibility Index (CSII) measure is computed for the low-level high-level ($CSII_{high}$), medium-level ($CSII_{mid}$), and ($CSII_{low}$) segments of each speech sentence, which can predict the intelligibility of peak-clipping and centering-clipping distortions in a speech signal [34]. Besides, the Normalized Covariance Metric (NCM) [34] and the Short-Time Objective Intelligibility (STOI) [50] are also conducted for intelligibility evaluation of enhanced speech.

3.5 Results and Discussions

In this section, we introduce two experimental results along with the respective discussion on the achieved results. The first experiment is performed on the 28-speaker database. We conduct a series of ablation experiments to evaluate our method. In the second experiment, we evaluate our method and several reproducible methods on different sizes of training data.

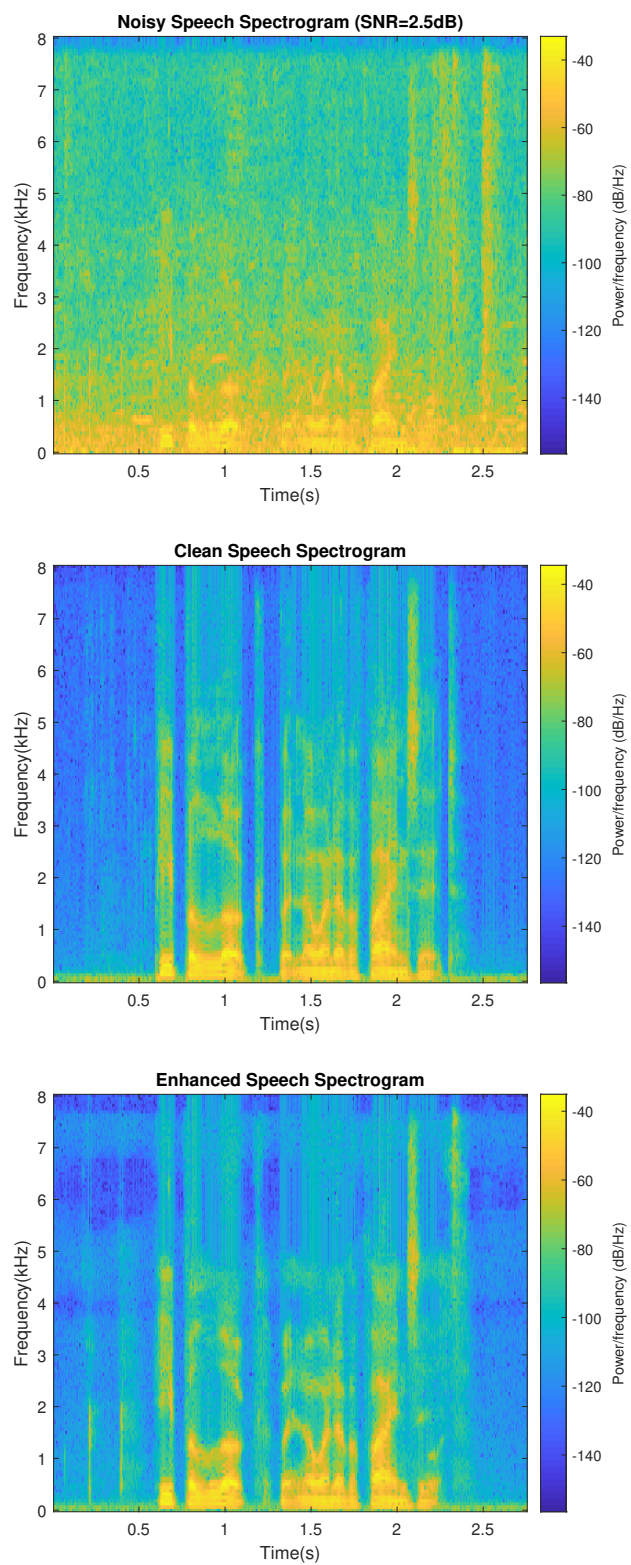


FIGURE 3.5: Spectrograms of selected utterance (SNR=2.5dB) enhanced with our method.

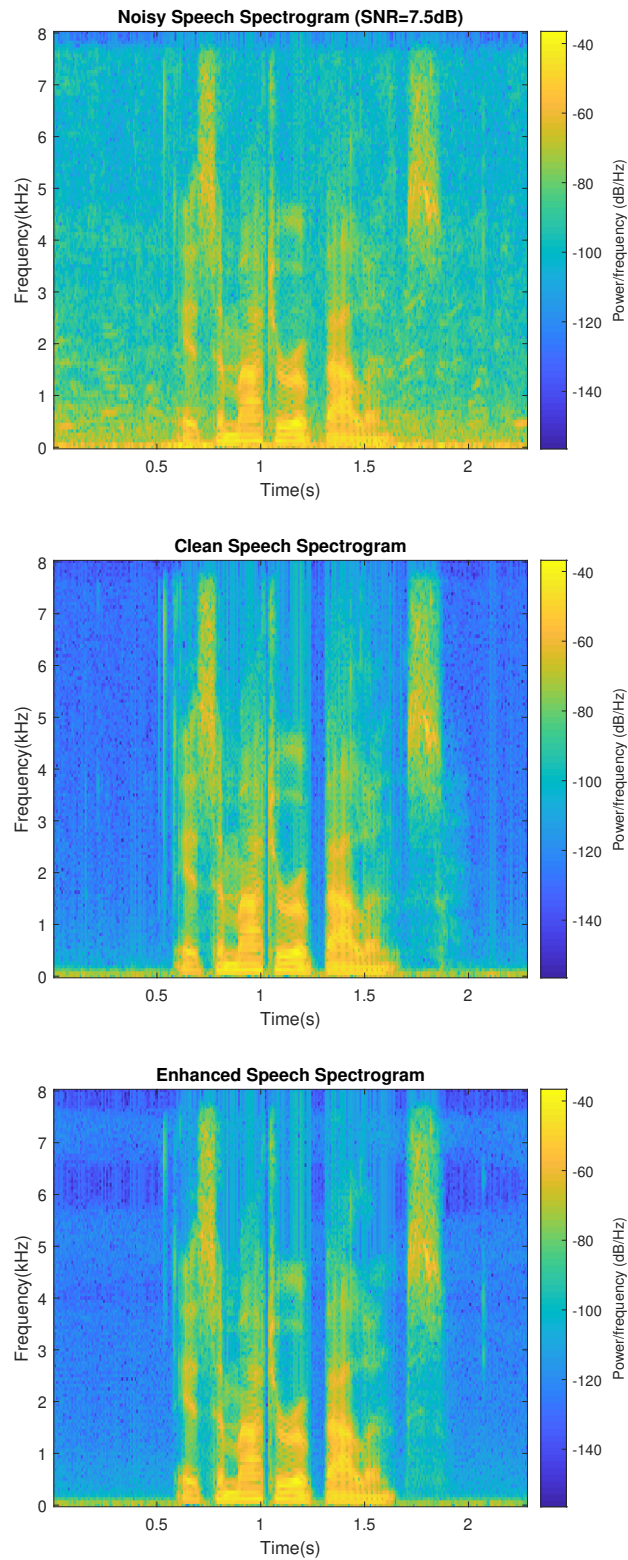


FIGURE 3.6: Spectrograms of selected utterance (SNR=7.5dB) enhanced with our method.

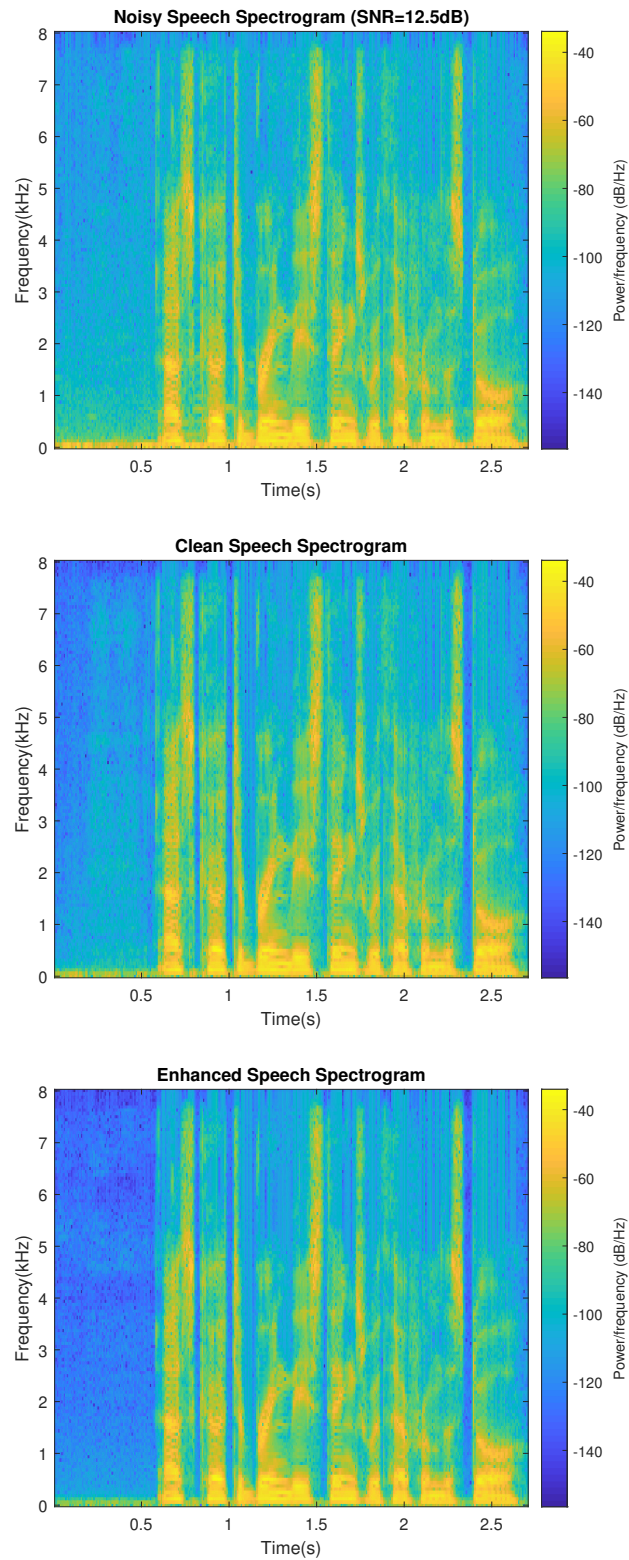


FIGURE 3.7: Spectrograms of selected utterance (SNR=12.5dB) enhanced with our method.

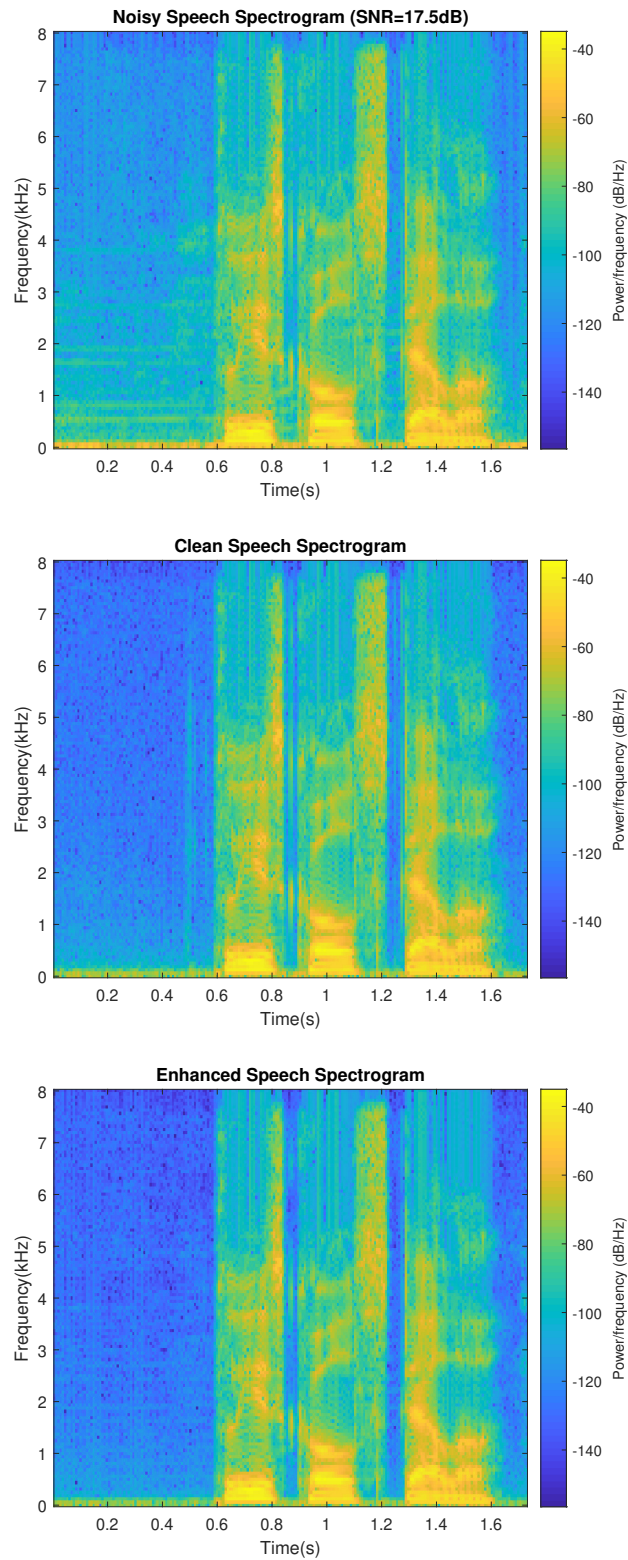


FIGURE 3.8: Spectrograms of selected utterance (SNR=17.5dB) enhanced with our method.

The performances of our proposed method and several compared methods on the 28-speaker database are listed in Table 3.1. Along with the previous evaluation strategy, we report the experimental results in terms of several main evaluation metrics including PESQ, CSIG, CBAK, CVOL, SSNR, and STOI.

In our ablation experiment, we activate the networks P and Q with corresponding loss functions by controlling the parameter λ for inverse mapping learning from output space to input space. We adopt SERGAN architecture in this chapter. Thus, the experimental results are the same as SERGAN \dagger if we set $\lambda_1 = 0, \lambda_2 = 0$. We just activate the network P and learn the inverse mapping from the generated space to the latent space when we set $\lambda_1 = 1, \lambda_2 = 0$. Compared with the original architecture (SERGAN) as we can see in Table 3.1, the experimental results show that our method achieves higher evaluation scores in terms of PESQ (2.52 to 2.57), CSIG (3.66 to 3.78), CBAK (3.18 to 3.23), CVOL (3.06 to 3.16), which relatively improves by 1.98%, 3.28%, 1.57%, and 5.23%, respectively. Moreover, we also find that the evaluation score decreases slightly in terms of SSNR (9.40 to 9.32) and remains the same in terms of STOI (0.937). We can infer that the proposed method of adversarial multi-task learning can further improve speech representation learning and speech enhancement performance. Compared with the previous methods, the proposed method can obtain competitive results based on 28-speaker training data.

Further, when we activate Q and inactivate P (i.e., $\lambda_1 = 0, \lambda_2 = 1$), the performance degrades slightly compared with the first ablation experiment (i.e., $\lambda_1 = 1, \lambda_2 = 0$) but still obtains a slight improvement compared with the original SERGAN architecture. We infer that re-excavating information from the input data domain by inverse mapping learning can improve representation learning and speech enhancement performance. However, the network P learning inverse mapping from the generated data domain to the latent domain is more effective in speech enhancement improvement than network Q .

Naturally, when we activate P and Q simultaneously (i.e., $\lambda_1 = 1, \lambda_2 = 1$), our method further improves the enhancement performance. Thus, we can infer that our proposed adversarial multi-task learning based method can improve speech representation learning and speech enhancement performance by inverse mapping learning. Moreover, when we add the multi-head self-attention [57] layer in the bottleneck layer of network G and Q , the evaluation results are further improved and obtain a competitive score compared with the state-of-the-art method.

To further demonstrate the effectiveness of our proposed method, we also unfold the details of the evaluation results with two reproducible methods (i.e., SEGAN and SERGAN) with a more comprehensive evaluation metric in Table 3.2. Compared with the two methods, we can find that our method improves speech enhancement performance in each SNR condition and evaluation metrics. Moreover, through careful comparison from high SNR to low SNR (17.5dB to 2.5dB), we find that our method performs better in lower SNR. In particular, the intelligibility improvement is dramatic. For example, in the 17.5dB scene, the NCM evaluation score of our method (0.997) is similar to other methods (0.994 and 0.997). However, in 2.5dB, the NCM score of our method (0.967) is much higher than other methods (0.928 and 0.959). Thus, we infer that our method performs well in low SNR, in terms of speech intelligibility and quality.

To visualize the performance, we also present the spectrograms of four selected speech utterances in different SNR in Figure 3.6 (2.5dB and 7.5dB) and Figure 3.8 (12.5dB and 17.5dB). From the top to bottom, the figures are noisy, clean, and enhanced speech waveforms. We can observe that our method can enhance noisy speech. Compared to Figure 3.6 and 3.8, it seems that the enhancement performance in low SNR (2.5dB and 7.5dB) is more effective than in high SNR (12.5dB and 17.5dB).

Moreover, we conduct experiments in different sizes of training data to further explore the effectiveness of our method. As we can see from Table 3.3, our method obtains further improvement in terms of speech quality and intelligibility with more training data. We also find that not all the evaluation metrics can achieve further improvement with more training data. We speculate that the results may be caused by the different distribution of training data.

3.6 Conclusions

In this chapter, we propose a novel adversarial multi-task learning with inverse mapping method for speech enhancement. Based on the generative adversarial structure, two extra networks are proposed to learn the inverse mappings from the generated distribution to the input data domain with the addition of two new functions (i.e. the latent loss and equilibrium loss functions). Working with the original adversarial loss function, our adversarial multi-task learning with inverse mapping method improves information capture and speech representation learning. The experimental results demonstrate that our proposed method can greatly improve speech enhancement performance in terms of speech quality and intelligibility, especially in a low SNR scene. Moreover, the multi-head self-attention is also effective to locate specific information and learn long-range dependencies of the speech signals, and improve speech enhancement further.

References

- [1] MA Abd El-Fattah, Moawad Ibrahim Dessouky, Salah M Diab, and Fathi El-Sayed Abd El-Samie. Speech enhancement using an adaptive Wiener filtering

- approach. *Progress in Electromagnetics Research*, 4:167–184, 2008. doi: <https://doi.org/10.2528/PIERM08061206>.
- [2] Deepak Baby and Sarah Verhulst. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 106–110, 2019. doi: <https://doi.org/10.1109/ICASSP.2019.8683799>.
- [3] Jacob Benesty, Jesper Rindom Jensen, Mads Graesboll Christensen, and Jingdong Chen. *Speech enhancement: A signal subspace perspective*. Academic Press, 1 edition, 2014. doi: <https://doi.org/10.1016/B978-0-12-800139-4.00009-8>.
- [4] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. doi: <https://doi.org/10.1023/A:1007379606734>.
- [5] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2172–2180, 2016.
- [6] Zhe Chen, Ruili Wang, Zhen Zhang, Huibin Wang, and Lizhong Xu. Background–foreground interaction for moving object detection in dynamic scenes. *Information Sciences*, 483:65–81, 2019. doi: <https://doi.org/10.1016/j.ins.2018.12.047>.
- [7] Fu-Kai Chuang, Syu-Siang Wang, Jieh-weih Hung, Yu Tsao, and Shih-Hau Fang. Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3173–3177, 2019. doi: <https://doi.org/10.21437/Interspeech.2019-2108>.
- [8] Feng Deng, T. Jiang, Xiaorui Wang, Chen Zhang, and Y. Li. NAAGN: noise-aware attention-gated network for speech enhancement. In *Proceedings of the*

- 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2457–2461, 2020. doi: <http://doi.org/10.21437/Interspeech.2020-1133>.
- [9] Mohamed Djendi and Rédha Bendoumia. Improved subband-forward algorithm for acoustic noise reduction and speech quality enhancement. *Applied Soft Computing*, 42:132–143, 2016. doi: <https://doi.org/10.1016/j.asoc.2016.01.049>.
- [10] Chris Donahue, Bo Li, and Rohit Prabhavalkar. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5024–5028. IEEE, 2018. doi: <https://doi.org/10.1109/ICASSP.2018.8462581>.
- [11] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, pages 1–18, 2017.
- [12] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 658–666, 2016.
- [13] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, pages 1–18, 2017.
- [14] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712. IEEE, 2015. doi: <http://doi.org/10.1109/ICASSP.2015.7178061>.
- [15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky.

- Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. doi: https://doi.org/10.1007/978-3-319-58347-1_10.
- [16] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee. Densely connected progressive learning for LSTM-based speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5054–5058. IEEE, 2018. doi: <http://doi.org/10.1109/ICASSP.2018.8461861>.
- [17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. doi: <http://doi.org/10.1109/ICCV.2015.123>.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: <https://doi.org/10.1109/CVPR.2016.90>.
- [20] Feng Hou, Ruili Wang, Jun He, and Yi Zhou. Improving entity linking through semantic reinforced entity embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6843–6848. Association for Computational Linguistics, July 2020. doi: <http://doi.org/10.18653/v1/2020.acl-main.612>.
- [21] Tsun-An Hsieh, Hsin-Min Wang, Xugang Lu, and Yu Tsao. WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement. *IEEE Signal Processing Letters*, 27:2149–2153, 2020. doi: <http://doi.org/10.1109/SPLET.2020.3000000>.

[//doi.org/10.1109/LSP.2020.3040693](https://doi.org/10.1109/LSP.2020.3040693).

- [22] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(1):229–238, 2007. doi: <https://doi.org/10.1109/TASLP.2007.911054>.
- [23] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5077–5086, 2017. doi: <https://doi.org/10.1109/CVPR.2017.202>.
- [24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International conference on machine learning (ICML)*, pages 448–456, 2015.
- [25] Wanting Ji and Ruili Wang. A multi-instance multi-label dual learning approach for video captioning. *ACM Transactions on Multimedia Computing Communications and Applications*, 17(2s):1–18, 2021. doi: <https://doi.org/10.1145/3446792>.
- [26] Alexia Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard GAN. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, pages 1–26, 2019.
- [27] Mathew Shaji Kavalekalam, Jesper Kjar Nielsen, Jesper Bunsow Boldt, and Mads Grasboll Christensen. Model-based speech enhancement for intelligibility improvement in binaural hearing aids. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(1):99–113, 2019. doi: <https://doi.org/10.1109/TASLP.2018.2872128>.
- [28] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 1–17, 2018.

- [29] Xiaofei Li, Laurent Girin, Sharon Gannot, and Radu Horaud. Multichannel on-line dereverberation based on spectral magnitude inverse filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(9):1365–1377, 2019. doi: <https://doi.org/10.1109/TASLP.2019.2919183>.
- [30] Jae Lim and Alan Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 26(3):197–210, 1978. doi: <https://doi.org/10.1109/TASSP.1978.1163086>.
- [31] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–10, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: <http://doi.org/10.18653/v1/P17-1001>.
- [32] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Multi-task adversarial network for disentangled feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3743–3751, 2018. doi: <http://doi.org/10.1109/CVPR.2018.00394>.
- [33] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. 2013. ISBN 1466504218. doi: <https://doi.org/10.1201/9781420015836>.
- [34] Jianfen Ma, Yi Hu, and Philipos C Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387–3405, 2009. doi: <http://doi.org/10.1121/1.3097493>.
- [35] Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang (Fred) Juang. Adversarial feature-mapping for speech enhancement. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3259–3263, 2018. doi: <https://doi.org/10.21437/Interspeech.2018-2461>.

- [36] Zhong Meng, Jinyu Li, Yifan Gong, et al. Cycle-consistent speech enhancement. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1165–1169, 2018. doi: <http://doi.org/10.21437/Interspeech.2018-2409>.
- [37] Daniel Michelsanti and Zheng-Hua Tan. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2008–2012, 2017. doi: <https://doi.org/10.21437/Interspeech.2017-1620>.
- [38] Tomohiro Nakatani and Keisuke Kinoshita. A unified convolutional beamformer for simultaneous denoising and dereverberation. *IEEE Signal Processing Letters*, 26(6):903–907, 2019. doi: <https://doi.org/10.1109/LSP.2019.2911179>.
- [39] Aaron Nicolson and Kuldip K Paliwal. Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Communication*, 111:44–55, 2019. doi: <https://doi.org/10.1016/j.specom.2019.06.002>.
- [40] Zhiheng Ouyang, Hongjiang Yu, Wei-Ping Zhu, and Benoit Champagne. A fully convolutional neural network for complex spectrogram processing in speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5756–5760. IEEE, 2019. doi: <http://doi.org/10.1109/ICASSP.2019.8683423>.
- [41] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: speech enhancement generative adversarial network. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3642–3646, 2017. doi: <https://doi.org/10.21437/Interspeech.2017-1428>.
- [42] Santiago Pascual, Joan Serrà, and Antonio Bonafonte. Towards generalized speech enhancement with generative adversarial networks. In *Proceedings of*

- the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 1791–1795, 2019. doi: <https://doi.org/10.21437/Interspeech.2019-2688>.
- [43] Lukas Pfeifenberger, Matthias Zöhrer, and Franz Pernkopf. Eigenvector-based speech mask estimation for multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(12): 2162–2172, 2019. doi: <https://doi.org/10.1109/TASLP.2019.2941592>.
- [44] Yuanhang Qiu and Ruili Wang. Adversarial latent representation learning for speech enhancement. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2662 – 2666, 2020. doi: <https://doi.org/10.21437/Interspeech.2020-1593>.
- [45] Nasir Saleem and Muhammad Irfan Khattak. Multi-scale decomposition based supervised single channel deep speech enhancement. *Applied Soft Computing*, 95:1066666, 2020. doi: <https://doi.org/10.1016/j.asoc.2020.1066666>.
- [46] Suman Samui, Indrajit Chakrabarti, and Soumya K Ghosh. Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network. *Applied Soft Computing*, 74:583–602, 2019. doi: <https://doi.org/10.1016/j.asoc.2018.10.031>.
- [47] Pourya Shamsolmoali, Masoumeh Zareapoor, Ruili Wang, Deepak Kumar Jain, and Jie Yang. G-GANISR: Gradual generative adversarial network for image super resolution. *Neurocomputing*, 366:140–153, 2019. doi: <https://doi.org/10.1016/j.neucom.2019.07.094>.
- [48] Meet H Soni, Neil Shah, and Hemant A Patil. Time-frequency masking-based speech enhancement using generative adversarial network. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5039–5043. IEEE, 2018. doi: <http://doi.org/10.13140/RG.2.2.19312.15365>.
- [49] Gajan Suthokumar, Vidhyasaharan Sethu, Kaavya Sriskandaraja, and

- Eliathamby Ambikairajah. Adversarial multi-task learning for speaker normalization in replay detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6609–6613, 2020. doi: <http://doi.org/10.1109/ICASSP40776.2020.9054322>.
- [50] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 19(7):2125–2136, 2011. doi: <http://doi.org/10.1109/TASL.2011.2114881>.
- [51] Hassan Taherian, Zhong-Qiu Wang, Jorge Chang, and DeLiang Wang. Robust speaker recognition based on single-channel and multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 28:1293–1302, 2020. doi: <https://doi.org/10.1109/TASLP.2020.2986896>.
- [52] Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Proceedings of the 19th Conference of the International Speech Communication Association (INTERSPEECH)*, volume 2018, pages 3229–3233, 2018. doi: <https://doi.org/10.21437/Interspeech.2018-1405>.
- [53] Naohiro Tawara, Tetsunori Kobayashi, and Tetsuji Ogawa. Multi-channel speech enhancement using time-domain convolutional denoising autoencoder. In *Proceedings of 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 86–90, 2019. doi: <https://doi.org/10.21437/Interspeech.2019-3197>.
- [54] Yan-Hui Tu, Jun Du, and Chin-Hui Lee. Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(12):2080–2091, 2019. doi: <https://doi.org/10.1109/TASLP.2019.2940662>.
- [55] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi.

- Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In *Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW)*, pages 146–152, 2016. doi: <http://doi.org/10.21437/SSW.2016-24>.
- [56] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 352–356, 2016. doi: <https://doi.org/10.21437/Interspeech.2016-159>.
- [57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.
- [58] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *Proceedings of Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE, 2013. doi: <http://doi.org/10.1109/ICSDA.2013.6709856>.
- [59] Zhong-Qiu Wang and DeLiang Wang. Deep learning based target cancellation for speech dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 28:941–950, 2020. doi: <https://doi.org/10.1109/TASLP.2020.2975902>.
- [60] Donald S Williamson and DeLiang Wang. Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM transactions on audio, speech, and language processing (TASLP)*, 25(7):1492–1501, 2017. doi: <https://doi.org/10.1109/TASLP.2017.2696307>.
- [61] Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. Deep learning for environmentally robust

speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–28, 2018. doi: <https://doi.org/10.1145/3178115>.

- [62] Han Zhao, Shuayb Zarar, Ivan Tashev, and Chin-Hui Lee. Convolutional-recurrent neural networks for speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2401–2405. IEEE, 2018. doi: <http://doi.org/10.1109/ICASSP.2018.8462155>.
- [63] Hao Zheng, Ruili Wang, Wanting Ji, Ming Zong, Wai Keung Wong, Zhihui Lai, and Hexin Lv. Discriminative deep multi-task learning for facial expression recognition. *Information Sciences*, 533:60–71, 2020.
- [64] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11127–11135, 2019. doi: <http://doi.org/10.1109/CVPR.2019.01138>.

Chapter 4

Self-Supervised Learning Based Phone-Fortified Speech Enhancement

Recent speech enhancement methods focus on further optimization of network structures and hyperparameters, however, ignore inherent speech characteristics (e.g., phonetic characteristics), which are important for networks to learn and reconstruct speech information. In this chapter, we propose a novel Self-Supervised learning based Phone-Fortified (SSPF) method for speech enhancement. Our method explicitly imports phonetic characteristics into a deep complex convolutional network via a Contrastive Predictive Coding (CPC) model pre-trained with self-supervised learning. This operation can greatly improve speech representation learning and speech enhancement performance. Moreover, we also apply the self-attention mechanism to our model for learning long-range dependencies of a speech sequence, which further improves the performance of speech enhancement.

4.1 Introduction

Speech enhancement is an important speech processing task aiming to improve the intelligibility and overall perceptual quality of a contaminated speech signal [19]. The intelligibility is a measurement of how comprehensible a speech signal is, while the perceptual quality measures how easy it is for a listener to perceive the content of a speech signal. Normally, a perceptual high-quality speech sounds more natural, rhythmic, yet less raspy, hoarse or scratchy, etc [19].

In the real world, additive noise (e.g., fan noise) and convolutional noise (e.g., room reverberation) are two common noise types that can degrade speech signals drastically. Correspondingly, many approaches (e.g., denoising, dereverberation) have been proposed and widely used in practical applications such as mobile communication [16], hearing-aids [13], and noise-robust speech recognition [38]. However, the performance of speech enhancement, especially in a real-life environment, still needs to be improved further.

Currently, data-driven deep learning based methods have been thriving in speech signal processing [30, 34], computer vision [12, 28, 42], and natural language processing [8]. For speech enhancement, existing deep learning based models, which pursued the optimal structures with dozens of network layers, have improved the performance of speech enhancement in different scenarios [2, 5, 24, 32, 37]. However, the performance might not be improved all along if we just stack the network layers exhaustively [15]. In addition, ignoring essential information of speech signals (e.g., phase and phonetic information) is also a challenging issue in speech enhancement [35].

For phase-aware speech enhancement, deep complex network based methods have demonstrated their effectiveness in dealing with complex-valued spectrums [10]. A deep complex network was originally proposed to construct richer and more versatile

representations of an image or audio signal [36]. Based on this network, recently, Hsieh et al. [9] proposed a Phone-Fortified Perceptual loss (PFP) for enhancing network optimization. They indicated that the phonetic characteristic information is the key to optimizing speech enhancement with respect to human perception, but the latent features for speech characteristics learning in previous models seemed to be lacking in phonetic characteristic information. Moreover, they also indicated that the objective functions based on point-wise distances might not fully reflect the perceptual difference between noisy and clean speech signals. Thus, the phonetic characteristics extracted by a pre-trained Contrastive Predictive Coding (CPC) model were introduced in their model but just for loss calculation.

Inspired by [9], we propose a new speech enhancement method, which focuses on improving speech representation learning via importing the phonetic characteristics into an improved deep complex network explicitly. We adopt a self-supervised learning based CPC model for speech phonetic information extraction because of CPC’s great speech representation learning capability. To import phonetic characteristics, we propose a new feature embedding network to re-embed the extracted features and then fuse them with the original frequency spectrum features. We consider that explicitly supplementing speech phonetic information can effectively enhance speech representation learning. Moreover, we also apply the self-attention mechanism to the deep complex network specifically, which aids in learning long-range dependencies of a speech sequence and improving the performance of speech enhancement further. In our experiments, we explore multiple CPC-based pre-trained models for speech phonetic information extraction and compare their performance fully. We also investigate the impact of the size of training data on our enhancement model and unfold the results in terms of Signal-to-Noise Ratios (SNR) and noise types.

In the following, we give the related work in Section 4.2. We describe the details of our model in Section 4.3. In Section 4.4, we describe the setup of experiments. The

experimental results are presented in Section 4.5. Finally, we report our conclusions in Section 4.6.

4.2 Related Work

The existing work related to speech phase information preservation and representation learning is introduced in this section.

4.2.1 Phase Information Preservation

For phase information preservation, an end-to-end speech processing framework has been considered as a plausible solution, which receives a raw speech waveform as input and outputs the processed speech waveform directly [23]. Since a raw speech waveform naturally contains phase information, the end-to-end speech enhancement can preserve the phase information from the contaminated speech sequence without extra handcrafted feature pre-processing. Generally, the handcrafted pre-processing operation such as traditional speech feature extraction may only capture acoustic information, but ignore other important information such as the speech phase [44]. The end-to-end framework can alleviate this problem by taking in the raw speech waveform. However, for speech enhancement, reusing the phase information of noisy speech generally causes a serious mismatch between reconstructed speech and clean speech, especially under extremely noisy conditions [10].

Further, jointly estimating the speech magnitude and phase information with a complex-valued network is another approach [44]. Unlike the real-valued network that only changes the scale of the magnitude spectral mapping without the phase information processing [35], the complex-valued network [36] learns speech magnitude and phase information with the real and imaginary part, respectively, which

has been proven to be an effective framework [9, 10] for speech enhancement. Thus, we also adopt the complex-valued network for speech magnitude and phase response preservation in this research.

4.2.2 Speech Representation Learning

Learning appropriate speech representation is a fundamental and effective way to improve speech signal processing. With the development of speech signal processing, different speech feature representations were proposed such as Mel-Frequency Cepstral Coefficients (MFCCs), a general all-purpose frame-level acoustic feature [45]; Identity vector (I-vector), a high-dimensional utterance-level speech representation [4]; Speech2vec (i.e., speech version of word2vec [20]), a semantic representation of an audio segment with a fixed-length vector [3]. These speech feature representations have been used widely in specific speech tasks.

Recently, a self-supervised learning based CPC model was developed to extract useful data representation from high-dimensional data space with a powerful autoregressive model [22]. Specifically, the probabilistic contrastive loss was proposed to induce the latent space to capture information that was maximally useful to predict future samples. The self-supervised learning mechanism enables CPC to learn a general and effective representation with massive unlabelled data. CPC was tested in different data modalities such as speech, images, natural language and obtained promising results [22, 27, 43]. In this chapter, we introduce two CPC-based models (i.e., wav2vec [27] and vq-wav2vec [1]) for speech representation learning.

4.2.2.1 Wav2vec

Wav2vec [27], a pre-trained CPC model as shown in Figure 4.1, can learn a general speech representation by training with large amounts of unlabelled raw audio data.

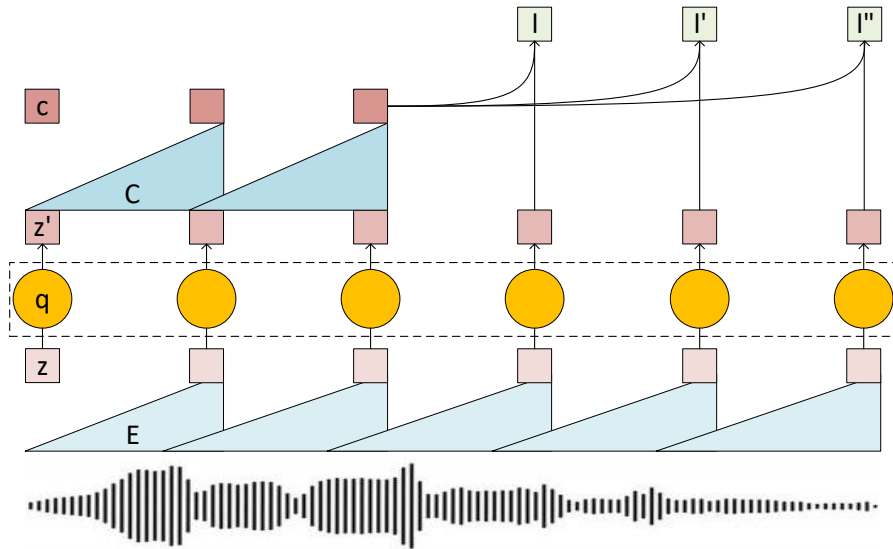


FIGURE 4.1: The framework of wav2vec and vq-wav2vec. For wav2vec, the Encoder (E) network maps the raw audio to a dense representation z . z is aggregated into a Context (C) network for representation c , which refers to the contrastive loss calculation (l) with the future samples. With the addition of a quantized (q) layer, this framework represents vq-wav2vec.

The model consists of two convolutional neural networks (i.e., an encoder network and a context network). The encoder network embeds the raw audio signal in a latent space and outputs a low-frequency representation to the context network (also known as an aggregator), which creates a contextualized vector representation by combining the latent representation from multiple time steps. The model can be trained to distinguish a future sample from the distractor samples, which is drawn from a proposal distribution, by minimizing the contrastive loss for each step. After training, the output of the context network can be considered as the desired representation of the input audio.

4.2.2.2 Vq-wav2vec

Vq-wav2vec [1] is based on wav2vec, as shown in Figure 4.1. It has an architecture like wav2vec but with an additional quantization module between the encoder network and the context network. The quantization module replaces the original representation z by z' from a fixed size codebook, of which the one-hot representation can be computed by using the Gumbel-Softmax or K-means clustering approaches [1]. Vq-wav2vec, which learns the discrete representations of fixed length segments of an audio signal, enables well-performing language processing algorithms [6] to be applied directly to speech data. In this chapter, we use both wav2vec and vq-wav2vec as phonetic characteristics extractors for speech enhancement model training and compare their performances in the Results Section.

4.3 The Proposed Method

In this chapter, we propose a new Self-Supervised learning based Phone-Fortified (SSPF) method for speech enhancement. Our method adopts a deep complex convolutional network to estimate a complex ratio mask for noisy information filtering. As shown in Figure 4.2, the deep complex network is a refined U-Net architecture [26] and incorporates multiple well-defined complex-valued blocks to deal with complex-valued spectrum [36]. In detail, the complex U-Net adopts multiple complex convolutional and transposed convolutional layers with skip-connections [21], complex batch normalization, and LeakyRelu activation [9] as the main components, which are admittedly functional parts to learn representations of multiple data modalities effectively such as image and speech. For speech enhancement, the complex-valued architecture with real and imaginary parts is used to learn speech magnitude and phase information simultaneously [36].

Referring to Figure 4.2, our speech enhancement model converts a noisy speech signal to an enhanced speech signal with a learnable complex mask derived from the complex U-Net model. To improve the speech representation learning of our speech enhancement model, we employ a pre-trained CPC-based model to extract the phonetic characteristics, which are then fused with the standard frequency spectrum feature converted by Short-Time Fourier Transform (STFT). Here, a simple feature embedding network is proposed to re-embed and normalize the representation of the phonetic characteristics so that it can be fused with the frequency spectrum feature by point-wise addition. The proposed feature embedding network consists of multiple transposed convolutional layers followed by Relu activation and max-pooling [29] and the network parameters are trained along with the complex U-Net simultaneously. We apply a self-attention layer in the middle of complex U-Net since the self-attention layer can learn the long-range dependencies of a speech sequence and further improve speech representation learning. At last, the frequency spectrum feature is point-wise multiplied with the complex-valued ratio mask again to derive the enhanced spectrum, and then the inverse STFT module transforms the enhanced spectrum to a speech waveform.

During the loss calculation, the same pre-trained CPC model is applied again to transform the waveform into a batch of sequence vectors, which are rich in phone-fortified information for a speech evaluation. We follow the effective phone-fortified perceptual loss proposed in [9]. The formulation can be described as below:

$$L_{pfp}(x, \hat{x}) := E_{x, \hat{x} \sim D}[\|\phi_{cpc}(x) - \phi_{cpc}(f(\hat{x}))\|_1], \quad (4.1)$$

where x denotes the clean speech; \hat{x} denotes the paired noisy speech; ϕ_{cpc} is the pre-trained CPC model for phonetic representation extraction; f denotes the enhancement procedure. The PFP loss calculates the absolute distance between a clean and an enhanced speech phonetic vector.

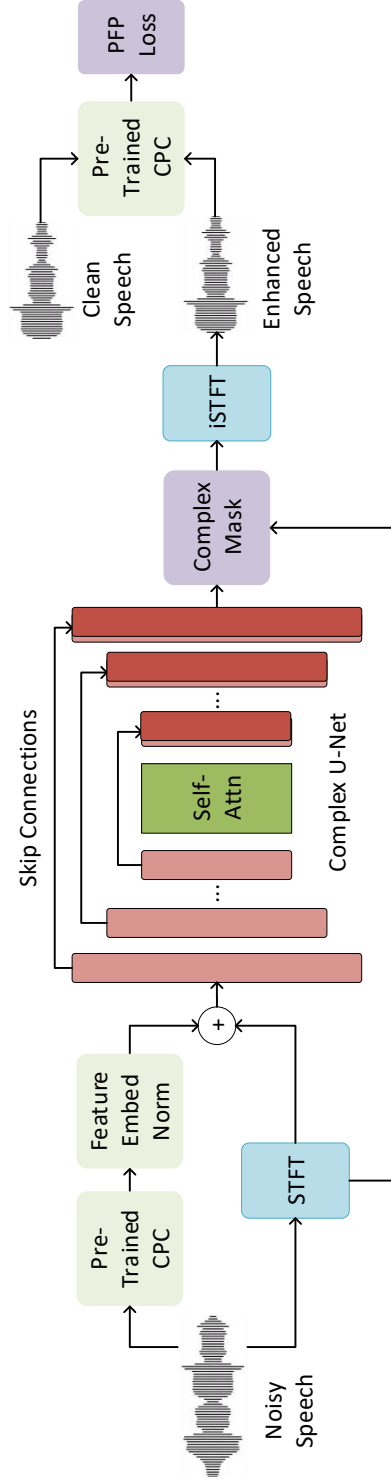


FIGURE 4.2: The framework of our proposed speech enhancement model based on a complex U-Net and CPC. The complex U-Net estimates a complex-valued ratio mask with the fused noisy speech representation. The mask can filter the noisy spectrum and achieve the enhanced spectrum with inverse STFT. The CPC-based pre-trained model extracts linguistic information of an enhanced waveform and paired clean waveform for PFP loss calculation.

4.4 Experiments

4.4.1 Database

The VCTK database [39, 40] is an open and standard speech corpus for performance evaluation of speech enhancement systems. The original clean speech was selected from the Voice Bank corpus [41]. In addition, eight real noise files and two artificially generated noise files were used to generate paired noisy speech.

For training, two sub-databases were created: one has 28 speakers, which includes 14 males and 14 females with the same accent (England) [40]; another one has 56 speakers, which includes 28 males and 28 females with different accents (Scotland and United States) [39]. The SNR values were set to 15dB, 10dB, 5dB and 0dB. Moreover, each clean speech waveform was normalized, and the silence segments were trimmed off at the beginning and the ending when the silence segments were longer than 200ms.

Another two speakers (a male and a female), not included in the training data, were selected as the test data with an England accent. Five other noise types, different from training data, were selected from the DEMAND database¹, including a domestic noise (in a living room), an office noise (in an office space), a transport noise (in a bus) and two street noises (in an open area cafeteria and a public square). The SNR values were set to 17.5dB, 12.5dB, 7.5dB and 2.5dB, respectively.

4.4.2 Setup

Before training, we randomly separate a validation part from the training data at a ratio of 9:1. The model is optimized using the RAdam optimizer [18] with a learning

¹<http://parole.loria.fr/DEMAND/>

TABLE 4.1: The evaluation results of various methods with the 28-speaker VCTK training data. The compared methods are Wiener filtering [17], ALRL [24], BiLSTM [7], CRN-MSN [34], AMTL-IM [25], NAAGN [5], U-NetC [2, 37], PHASEN [44], HiFi-GAN [31], T-GSA [14]. “-” denotes that the result is not reported or not available. “†” denotes that we reproduced the results with the provided open resource. The best scores are highlighted in bold.

Models	PESQ	CSIG	CBAK	CVOL	STOI
Noisy	1.97	3.35	2.44	2.63	0.921
Wiener	2.22	3.23	2.68	2.67	-
ALRL	2.57	3.78	3.23	3.16	0.937
BiLSTM	2.70	3.99	2.95	3.34	0.925
CRN-MSE	2.74	3.86	3.14	3.30	0.934
AMTL-IM	2.88	4.01	3.50	3.51	0.945
NAAGN	2.90	4.13	3.50	3.51	0.948
U-NetC	2.90	4.22	3.32	3.58	0.938
HiFi-GAN	2.94	4.07	3.07	3.49	-
PHASEN	2.99	4.21	3.55	3.62	-
T-GSA	3.06	4.18	3.59	3.62	-
W2V†	2.98	4.01	3.46	3.50	0.945
W2V _F	3.02	4.11	3.52	3.60	0.943
W2V _{FSA}	3.04	4.17	3.59	3.63	0.945
W2V-G	2.99	4.09	3.46	3.55	0.944
W2V-G _F	3.00	4.12	3.48	3.57	0.944
W2V-G _{FSA}	2.96	4.04	3.47	3.50	0.944
W2V-K	2.93	4.06	3.45	3.50	0.942
W2V-K _F	2.95	4.11	3.54	3.57	0.943
W2V-K _{FSA}	3.00	4.14	3.46	3.58	0.942

rate of 0.0001 and weight decay of 0.1. The model is trained for 100 epochs with a batch size of 8. For each epoch, we save the best model according to the performance evaluated with the validation data. During the inference stage, each noisy speech is point-wise multiplied with the complex mask for noisy information filtering and then is converted to an enhanced speech waveform.

4.4.3 Evaluation Metrics

Although subjective evaluation is accurate and reliable, it is costly and time-consuming [11]. Many objective evaluation measures can evaluate enhanced speech performance with high correlation. For speech quality evaluation, we use an effective full-reference speech quality evaluation algorithm [11] namely Perceptual Evaluation of Speech Quality (PESQ: from -0.5 to 4.5), which compares each sample of the reference signal (clean speech) to each corresponding sample of the degraded signal, and analyses sample-by-sample after a temporal alignment of corresponding excerpts of reference and testing signals. Moreover, we also implement the composite evaluation metrics of the enhanced speech including the predicted Mean Opinion Score (MOS) of signal distortion (CSIG: from 1 to 5), background noise distortion (CBAK: from 1 to 5), and overall quality (COVL: from 1 to 5). For speech intelligibility evaluation, we adopt the Short-Time Objective Intelligibility (STOI) [33] that is based on a correlation coefficient between the temporal envelopes of the clean and degraded speech.

4.5 Results

As shown in Table 4.1, we explore three CPC-based models for phonetic information extraction and loss calculation: Wav2Vec (W2V), vq-Wav2Vec with Gumbel-Softmax (W2V-G), and vq-Wav2Vec with K-means clustering (W2V-K). To implement the ablation experiments, we test our speech enhancement model with the frequency spectrum feature only (i.e., W2V, W2V-G, W2V-K as shown in Table 4.1), with the phonetic embedding fused feature (i.e., W2V_F, W2V-G_F, W2V-K_F), and further with self-attention mechanism (i.e., W2V_{FSA}, W2V-G_{FSA}, W2V-K_{FSA}).

TABLE 4.2: The evaluation results of W2V_{FS}A with 56-speaker and 84-speaker (the mixture of 28-speaker and 56-speaker). Meanwhile, all scenes with different SNR and noise types are demonstrated separately. The best scores are highlighted in bold.

Metric	W2V _{FS} A	2.5dB	7.5dB	12.5dB	17.5dB	Living	Office	Bus	Cafe	Square	Average
PESQ	56-spkr	2.61	3.04	3.29	3.55	2.86	3.57	3.52	2.61	2.97	3.10
	84-spkr	2.67	3.08	3.31	3.53	2.93	3.55	3.52	2.65	3.02	3.13
CSIG	56-spkr	3.75	4.22	4.47	4.70	4.02	4.70	4.63	3.82	4.20	4.27
	84-spkr	3.86	4.26	4.47	4.66	4.09	4.68	4.62	3.87	4.24	4.30
CBAK	56-spkr	3.12	3.50	3.73	3.99	3.38	3.98	3.74	3.25	3.53	3.57
	84-spkr	3.19	3.55	3.76	3.99	3.44	3.98	3.79	3.28	3.57	3.61
CVOL	56-spkr	3.17	3.64	3.90	4.16	3.44	4.17	4.10	3.21	3.59	3.70
	84-spkr	3.26	3.67	3.90	4.11	3.51	4.14	4.09	3.26	3.64	3.72
STOI	56-spkr	0.920	0.950	0.960	0.966	0.939	0.967	0.964	0.924	0.946	0.947
	84-spkr	0.923	0.952	0.961	0.967	0.943	0.967	0.965	0.927	0.948	0.950

Compared with the previous methods, $W2V_{FSA}$ (i.e., wav2vec model with fused features and self-attention mechanism) obtains the best score in CBAK (3.59), and CVOL (3.63). In terms of other metrics, $W2V_{FSA}$ also achieves competitive results. Focusing on the ablation experiments, we find that the performance can be improved gradually when we import the phonetic information and employ the self-attention mechanism with both wav2vec and vq-wav2vec. Thus, we conclude that our proposed method can effectively learn speech representation and the phone-fortified method has a huge potential to improve speech enhancement. In addition, we find that the wav2vec based models outperform vq-wav2vec models. We infer that the non-discrete representations learned by wav2vec outperform the discrete representations learned by vq-wav2vec for speech enhancement.

To further explore the effectiveness of our proposed method, we conduct another experiment with different sizes of training data. Meanwhile, we reveal the results in four SNR and five noise type scenes respectively as shown in Table 4.2. In this experiment, we present the evaluation results on the best $W2V_{FSA}$ model trained with 56-speaker and 84-speaker (the mixture of 28-speaker and 56-speaker). As we can see, with more training data, the $W2V_{FSA}$ model further improves the performance and achieves state-of-the-art performance in most scenes. Moreover, we also find that our method can perform better in scenarios where SNR is as low as 2.5dB.

4.6 Conclusions

In this chapter, we propose a novel Self-Supervised learning based Phone-Fortified method (SSPF) for speech enhancement. Our SSPF method can effectively estimate a complex ratio mask for noisy speech filtering with a self-attention mechanism boosted complex U-Net model. SSPF explicitly imports the phonetic characteristics

into the enhancement model via a self-supervised learning based CPC model to further improve speech phase estimation and representation learning. The experimental results demonstrate that our method outperforms previous methods in most evaluation metrics and achieves state-of-the-art performance with more training data in terms of speech quality and intelligibility.

References

- [1] Alexei Baevski, Steffen Schneider, and Michael Auli. Vq-wav2vec: Self-supervised learning of discrete speech representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.
- [2] Ahmet E Bulut and Kazuhito Koishida. Low-latency single channel speech enhancement using U-Net convolutional neural networks. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6214–6218. IEEE, 2020. doi: <https://doi.org/10.1109/ICASSP40776.2020.9054563>.
- [3] Yu-An Chung and James Glass. Speech2vec: a sequence-to-sequence framework for learning word embeddings from speech. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 811–815, 2018. doi: <https://doi.org/10.21437/Interspeech.2018-2341>.
- [4] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 19(4):788–798, 2010. doi: <https://doi.org/10.1109/TASL.2010.2064307>.
- [5] Feng Deng, T. Jiang, Xiaorui Wang, Chen Zhang, and Y. Li. NAAGN: noise-aware attention-gated network for speech enhancement. In *Proceedings of the*

- 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2457–2461, 2020. doi: <http://doi.org/10.21437/Interspeech.2020-1133>.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics, June 2019. doi: <http://doi.org/10.18653/v1/N19-1423>.
- [7] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712. IEEE, 2015. doi: <http://doi.org/10.1109/ICASSP.2015.7178061>.
- [8] Feng Hou, Ruili Wang, Jun He, and Yi Zhou. Improving entity linking through semantic reinforced entity embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6843–6848. Association for Computational Linguistics, 2020. doi: <http://doi.org/10.18653/v1/2020.acl-main.612>.
- [9] Tsun-An Hsieh, Cheng Yu, Szu-Wei Fu, Xugang Lu, and Yu Tsao. Improving perceptual quality by phone-fortified perceptual loss for speech enhancement. *arXiv preprint arXiv:2010.15174*, 2020.
- [10] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2472–2476, 2020. doi: <https://doi.org/10.21437/Interspeech.2020-2537>.

-
- [11] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(1):229–238, 2007. doi: <https://doi.org/10.1109/TASLP.2007.911054>.
- [12] Wanting Ji, Ruili Wang, Yan Tian, and Xun Wang. An attention based dual learning approach for video captioning. *Applied Soft Computing*, page 108332, 2021. doi: <https://doi.org/10.1016/j.asoc.2021.108332>.
- [13] Mathew Shaji Kavalekalam, Jesper Kjar Nielsen, Jesper Bunsow Boldt, and Mads Græsbøll Christensen. Model-based speech enhancement for intelligibility improvement in binaural hearing aids. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(1):99–113, 2019. doi: <https://doi.org/10.1109/TASLP.2018.2872128>.
- [14] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. T-GSA: transformer with gaussian-weighted self-attention for speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6649–6653. IEEE, 2020. doi: <http://doi.org/10.1109/ICASSP40776.2020.9053591>.
- [15] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 6391–6401, 2018.
- [16] Junfeng Li, Shuichi Sakamoto, Satoshi Hongo, Masato Akagi, and Yôiti Suzuki. Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication. *Speech Communication*, 53(5):677–689, 2011. doi: <https://doi.org/10.1016/j.specom.2010.04.009>.
- [17] Jae Lim and Alan Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 26(3):197–210, 1978. doi: <https://doi.org/10.1109/TASSP.1978.1163086>.

- [18] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the 8rd International Conference on Learning Representations (ICLR)*, 2019.
- [19] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. 2013. ISBN 1466504218. doi: <https://doi.org/10.1201/9781420015836>.
- [20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 2, pages 3111–3119. Curran Associates Inc., 2013.
- [21] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet a generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- [22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [23] Santiago Pascual, Antonio Bonafonte, and Joan Serra. SEGAN: speech enhancement generative adversarial network. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3642–3646, 2017. doi: <https://doi.org/10.21437/Interspeech.2017-1428>.
- [24] Yuanhang Qiu and Ruili Wang. Adversarial latent representation learning for speech enhancement. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2662–2666, 2020. doi: <https://doi.org/10.21437/Interspeech.2020-1593>.
- [25] Yuanhang Qiu, Ruili Wang, Feng Hou, Satwinder Singh, Zhizhong Ma, and Xiaoyun jia. Adversarial multi-task learning with inverse mapping for speech

- enhancement. *Under review of Applied Soft Computing*, 2021.
- [26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. doi: http://doi.org/10.1007/978-3-319-24574-4_28.
- [27] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. Wav2vec: unsupervised pre-training for speech recognition. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3465–3469, 2019. doi: <http://doi.org/10.21437/Interspeech.2019-1873>.
- [28] Pourya Shamsolmoali, Masoumeh Zareapoor, Ruili Wang, Deepak Kumar Jain, and Jie Yang. G-GANISR: Gradual generative adversarial network for image super resolution. *Neurocomputing*, 366:140–153, 2019. doi: <https://doi.org/10.1016/j.neucom.2019.07.094>.
- [29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations ICLR*, 2015.
- [30] Satwinder Singh, Ruili Wang, and Yuanhang Qiu. DEEPF0: end-to-end fundamental frequency estimation for music and speech signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65, 2021. doi: <https://doi.org/10.1109/ICASSP39728.2021.9414050>.
- [31] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. HiFi-GAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks. In *Proceedings of the 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4506–4510, 2020. doi: <http://doi.org/10.21437/Interspeech.2020-2143>.

- [32] Lei Sun, Jun Du, Li-Rong Dai, and Chin-Hui Lee. Multiple-target deep learning for LSTM-RNN based speech enhancement. In *Proceedings of the Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pages 136–140. IEEE, 2017. doi: <https://doi.org/10.1109/TASLP.2017.2746264>.
- [33] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 19(7):2125–2136, 2011. doi: <http://doi.org/10.1109/TASL.2011.2114881>.
- [34] Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Proceedings of the 19th Conference of the International Speech Communication Association (INTERSPEECH)*, volume 2018, pages 3229–3233, 2018. doi: <https://doi.org/10.21437/Interspeech.2018-1405>.
- [35] Ke Tan and DeLiang Wang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 28:380–390, 2019. doi: <https://doi.org/10.1109/TASLP.2019.2955276>.
- [36] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.
- [37] Dung N Tran and Kazuhito Koishida. Single-channel speech enhancement by subspace affinity minimization. In *Proceedings of the 21st Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2447–2451, 2020. doi: <https://doi.org/10.21437/Interspeech.2020-2982>.
- [38] Yan-Hui Tu, Jun Du, and Chin-Hui Lee. Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(12):2080–2091, 2019. doi: <https://doi.org/>

- 10.1109/TASLP.2019.2940662.
- [39] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In *Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW)*, pages 146–152, 2016. doi: <http://doi.org/10.21437/SSW.2016-24>.
- [40] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 352–356, 2016. doi: <https://doi.org/10.21437/Interspeech.2016-159>.
- [41] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *Proceedings of Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE, 2013. doi: <http://doi.org/10.1109/ICSDA.2013.6709856>.
- [42] Lei Wang, Xiaoguang Yuan, Ming Zong, Yujun Ma, Wanting Ji, Mingzhe Liu, and Ruili Wang. Multi-cue based four-stream 3d resnets for video-based action recognition. *Information Sciences*, 575:654–665, 2021. doi: <https://doi.org/10.1016/j.ins.2021.07.079>.
- [43] Anne Wu, Changhan Wang, Juan Pino, and Jiatao Gu. Self-supervised representations improve end-to-end speech translation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1491–1495, 2020. doi: <http://doi.org/10.21437/Interspeech.2020-3094>.
- [44] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. PHASEN: a phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 9458–9465, 2020. doi: <https://doi.org/10.1609/aaai.v34i05.6489>.

- [45] Andr'as Zolnay, Ralf Schluter, and Hermann Ney. Acoustic feature combination for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 457–460. IEEE, 2005. doi: <https://doi.org/10.1109/ICASSP.2005.1415149>.

Chapter 5

Summary

This chapter provides a summary of this thesis. Firstly, we give a research summary of contributions in Section 5.1, including latent space information exploration based on GAN-based architecture (Chapter 2); speech representation learning with adversarial multi-task learning method (Chapter 3); speech phase information and phonetic characteristics learning with self-supervised learning method (Chapter 4). Furthermore, we also discuss future work of speech enhancement in Section 5.2.

5.1 Research Summary

In this thesis, we proposed three novel methods for speech enhancement aiming to improve speech representation learning and enhancement performance. A recap of our methods and contributions is listed as follows:

- Chapter 2 presents a novel Adversarial Latent Representation Learning (ALRL) method for speech enhancement [8]. Based on adversarial feature learning,

ALRL employs an extra encoder to learn an inverse mapping from the generated data distribution to the latent space. The encoder establishes an inner connection with the generator and provides relevant latent information for adversarial feature modelling. A new loss function is proposed to implement the encoder mapping. In addition, the multi-head self-attention is also applied to the encoder for learning of long-range dependencies and further effective adversarial representations. The experimental results demonstrate that ALRL outperforms current GAN-based speech enhancement methods.

- Chapter 3 presents an adversarial multi-task learning with inverse mapping method for speech enhancement [9]. This method focuses on enhancing the generator’s capability of speech information capture and representation learning. To implement our method, two extra networks (namely P and Q) are developed to establish the inverse mapping from the generated distribution to the input data domains. Correspondingly, two new loss functions (i.e., latent loss and equilibrium loss) are proposed for the inverse mapping learning and the enhancement model training based on the original adversarial loss. The experimental results demonstrate that this method can effectively improve speech representation learning and outperform current methods in terms of speech quality and intelligibility.
- Chapter 4 presents a Self-Supervised learning based Phone-Fortified (SSPF) method for speech enhancement [10]. This method explicitly incorporates phonetic characteristics into a deep complex convolutional network via a Contrastive Predictive Coding (CPC) model pre-trained with self-supervised learning. This operation can greatly improve speech representation learning and speech enhancement performance. Moreover, we also apply the self-attention mechanism to this model for learning long-range dependencies of a speech sequence, which further improves the performance of speech enhancement. The

experimental results demonstrate that our SSPF method outperforms existing methods and achieves state-of-the-art performance in terms of speech quality and intelligibility.

5.2 Future Work

In this section, we propose some future work for speech enhancement research.

- **Enlarge experiment scale.** We will further conduct experiments with sufficient speech data including more realistic scenarios to evaluate the effectiveness and robustness of our methods. The training data should consider as many scenarios as possible to reflect the realistic environments and improve the adaptability of the speech enhancement model.
- **Multiple features fusion for speech representation learning.** Multiple features fusion can provide multiple hierarchies data representation for model training and mapping learning. In many research areas, feature fusion methods are used to achieve a more robust and effective model [2, 4, 12, 15]. Thus, further exploration about multiple features fusion in speech enhancement will be one of our future projects.
- **Novel neural networks for speech enhancement.** Recently, several novel architectures were proposed and made a breakthrough in many research areas such as attention based transformer architecture [14] and its variants [5, 6, 11]. Those models, adopting a revolutionary concept by eliminating recurrent or convolutional portions to improve information learning and result inference, will be applied to speech enhancement in our future work.

- **Applications of speech enhancement.** For robust speech recognition [3, 16], speaker recognition [7, 13], and speech synthesis [1], speech enhancement can be considered as front-end preprocessing and be used to improve the performance of those back-end applications. We will apply the proposed methods to back-end applications in future work.

References

- [1] Nagaraj Adiga, Yannis Pantazis, Vassilis Tsiaras, and Yannis Stylianou. Speech enhancement for noise-robust speech synthesis using wasserstein gan. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1821–1825, 2019. doi: <https://doi.org/10.21437/Interspeech.2019-2648>.
- [2] Khursheed Aurangzeb, Irfan Haider, Muhammad Attique Khan, Tanzila Saba, Kashif Javed, Tassawar Iqbal, Amjad Rehman, Hashim Ali, and Muhammad Shahzad Sarfraz. Human behavior analysis based on multi-types features fusion and von nauman entropy based features reduction. *Journal of Medical Imaging and Health Informatics*, 9(4):662–669, 2019.
- [3] Chris Donahue, Bo Li, and Rohit Prabhavalkar. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5024–5028. IEEE, 2018. doi: <https://doi.org/10.1109/ICASSP.2018.8462581>.
- [4] Yongming Huang, Kexin Tian, Ao Wu, and Guobao Zhang. Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. *Journal of Ambient Intelligence and Humanized Computing*, 10(5):1787–1798, 2019. doi: <https://doi.org/10.1007/s12652-017-0644-8>.

-
- [5] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
- [6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.
- [7] Ondřej Novotný, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, et al. Analysis of DNN speech signal enhancement for robust speaker recognition. *Computer Speech & Language*, 58:403–421, 2019. doi: <https://doi.org/10.1016/j.csl.2019.06.004>.
- [8] Yuanhang Qiu and Ruili Wang. Adversarial latent representation learning for speech enhancement. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2662 – 2666, 2020. doi: <https://doi.org/10.21437/Interspeech.2020-1593>.
- [9] Yuanhang Qiu, Ruili Wang, Feng Hou, Satwinder Singh, Zhizhong Ma, and Xiaoyun jia. Adversarial multi-task learning with inverse mapping for speech enhancement. *Under review of Applied Soft Computing*, 2021.
- [10] Yuanhang Qiu, Ruili Wang, Satwinder Singh, Zhizhong Ma, and Feng Hou. Self-supervised learning based phone-fortified speech enhancement. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 211 – 215, 2021. doi: <https://doi.org/10.21437/Interspeech.2021-734>.
- [11] David So, Quoc Le, and Chen Liang. The evolved transformer. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5877–5886. PMLR, 2019.
- [12] Linhui Sun, Jia Chen, Keli Xie, and Ting Gu. Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition. *International Journal of Speech Technology*, 21(4):931–940, 2018. doi: <https://doi.org/10.1007/s10261-018-0441-1>.

[//doi.org/10.1007/s10772-018-9551-4](https://doi.org/10.1007/s10772-018-9551-4).

- [13] Hassan Taherian, Zhong-Qiu Wang, Jorge Chang, and DeLiang Wang. Robust speaker recognition based on single-channel and multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 28:1293–1302, 2020. doi: <https://doi.org/10.1109/TASLP.2020.2986896>.
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [15] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–284, 2018.
- [16] Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–28, 2018. doi: <https://doi.org/10.1145/3178115>.


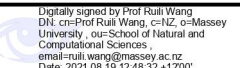
Appendix A

Statement of Contribution

I confirm that the “Statement of Contribution to Doctoral Thesis Containing Publications (DRC16)”, have been completed for each published article within the thesis, and are bound into the thesis and included in the electronic copy.

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS



We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yuanhang Qiu
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 2
<p>Please select one of the following three options:</p> <p><input checked="" type="radio"/> The manuscript/published work is published or in press</p> <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: Yuanhang Qiu and Ruili Wang*. Adversarial Latent Representation Learning for Speech Enhancement. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH), Virtual Event, Shanghai, China, 25-29 October 2020. pages 2662 - 2666. DOI: 10.21437/Interspeech.2020-1593 <p><input type="radio"/> The manuscript is currently under review for publication – please indicate:</p> <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: <p><input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal</p>	
Candidate's Signature:	Yuanhang Qiu 
Date:	15-Aug-2021
Primary Supervisor's Signature:	Prof Ruili Wang 
Date:	19-Aug-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/publication or collected as an appendix at the end of the thesis.

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS


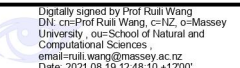
We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yuanhang Qiu
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 3
<p>Please select one of the following three options:</p> <p><input type="radio"/> The manuscript/published work is published or in press</p> <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: <p><input checked="" type="radio"/> The manuscript is currently under review for publication – please indicate:</p> <ul style="list-style-type: none"> • The name of the journal: Applied Soft Computing • The percentage of the manuscript/published work that was contributed by the candidate: 75.00 • Describe the contribution that the candidate has made to the manuscript/published work: <ul style="list-style-type: none"> - proposed an idea of adversarial multi-task learning based speech enhancement method. - implemented the experiments of speech intelligibility and quality evaluation <p><input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal</p>	
Candidate's Signature:	Yuanhang Qiu 
Date:	15-Aug-2021
Primary Supervisor's Signature:	Prof Ruili Wang 
Date:	19-Aug-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/publication or collected as an appendix at the end of the thesis.

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yuanhang Qiu
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 4
<p>Please select one of the following three options:</p> <p><input checked="" type="radio"/> The manuscript/published work is published or in press</p> <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: Yuanhang Qiu, Ruili Wang*, Satwinder Singh, Zhizhong Ma and Feng Hou. Self-Supervised Learning Based Phone-Fortified Speech Enhancement. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH), Hybrid Event, Brno, Czechia, 30 August-3 September 2021. pages 211 - 215. DOI:10.21437/Interspeech.2021-734. <p><input type="radio"/> The manuscript is currently under review for publication – please indicate:</p> <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: <p><input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal</p>	
Candidate's Signature:	Yuanhang Qiu 
Date:	15-Aug-2021
Primary Supervisor's Signature:	Prof Ruili Wang 
Date:	19-Aug-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/publication or collected as an appendix at the end of the thesis.