Mismatch Problem in Deep-learning Based Speech Enhancement

Hou Nana

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2022

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

13/12/2021

.....

Date

.....

Hou Nana

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

20/12/2021

.....

Date

.....

A/Prof. Chng Eng Siong

Authorship Attribution Statement

This thesis contains material from four paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as
(1) Hou, Nana, Chenglin Xu, Eng Siong Chng, and Haizhou Li. "Domain Adversarial Training For Speech Enhancement." In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2019.
(2) Hou, Nana, Chenglin Xu, Eng Siong Chng, and Haizhou Li. "Learning Disentangled Feature Representations for Speech Enhancement Via Adversarial Training." In ICASSP - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 666-670. IEEE, 2021.

The contributions of the co-authors are as follows:

- Prof. Chng and Prof. Li provided the initial idea direction and assisted to edit the final manuscript drafts for all papers.
- I proposed a potential solution, prepared the experimental setup, conducted all experiments, and prepared the initial manuscript drafts for all papers.
- Dr. Xu assisted to edit the final manuscript drafts for two papers.

Chapter 4 is published as Hou, Nana, Chenglin Xu, Joey Tianyi Zhou, Eng Siong Chng, and Haizhou Li. "Multi-Task Learning for End-to-End Noise-Robust Bandwidth Extension." In INTERSPEECH, pp. 4069-4073. 2020.

The contributions of the co-authors are as follows:

- Prof. Chng and Prof. Li suggested the idea direction and edited the final manuscript draft.
- I initialized a potential solution, detailed the experimental setup, conducted all experiments, and prepared the initial manuscript draft for the paper.
- Dr. Xu and Dr. Zhou polished the manuscript.

Chapter 5 is published as Hou, Nana, Eng Siong Chng, and Haizhou Li. " Two-step Hybrid Filterbanks Design for Time-domain Speech Enhancement." Will be submitted to TASLP 2021.

The contributions of the co-authors are as follows:

- Prof. Chng and Prof. Li provided invaluable suggestions and polished the final manuscript drafts.

- I proposed the novel idea, conducted all experiments, and prepared the manuscript.

13/12/2021

Hou Nana

Date

Hou Nana

Acknowledgements

I would like to thank my supervisors Prof. Chng Eng Siong and Prof. Li Haizhou for their invaluable supervision, support and tutelage during the course of my PhD degree. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. It is their generous help and support that have made my study and life in Singapore a wonderful time.

I also appreciate my colleagues and labmates Dr. Xu Chenglin, Dr. Xu Haihua, Mr. Lim Zhihao, Dr. Pham Van Tung, Dr. Rao Wei, Dr. Tian Xiaohai, Mr. Ma Duo, Ms. Ho Thi Nga, Ms. Vu Thi Ly, Mr. Kyaw Zin Tun, Mr. Chen Chen and Mr. Hu Yuchen at MICL labs for their support.

Additionally, I would like to express gratitude to my good friends Guo Mingrui, Cheng Suxia, Guo Wanhua, and Zhou Yi for their consistent encouragement.

My appreciation finally goes out to boyfriend Dr. Zhang Huaizheng and my parents for their endless love, which helps me endure dark times.

"I am a slow walker, but I never walk backwards."

— Abraham, Lincoln

To my dear family

Abstract

Speech enhancement aims to suppress background noise in noisy speech signals in order to improve speech perceptual quality and intelligibility. For tasks utilizing deep learning mechanisms, the training and testing data are usually assumed to have the same probability distribution. However, real-life scenarios often fail to meet this assumption. As a result, speech enhancement performance may degrade significantly, when faced with mismatched probability distributions between training and testing data. This thesis focuses on alleviating the problem of mismatched probability distributions for speech enhancement.

The mismatch problem in speech enhancement is caused by various factors, but in this work, we only focus on the following three scenarios: unseen noises in test data, missing high-frequency information under radio-channel testing condition (channel effect), and sensitive time-domain encoder/decoder. Specifically, we will clarify three factors, analysis impacts on speech enhancement, and propose three methods to solve this problem.

The first proposed method addresses the mismatch problem caused by the unseen noises in test data, under conditions with/without target-domain data. Specifically, we utilize the domain adversarial training (DAT) technique for domain transfer. If we have sufficient noisy target-domain data, a domain discriminator is proposed to learn general features with DAT in order to overcome domain mismatch problem. If we have no target-domain data, we will utilize the noise labels of the source-domain data to generate noise-agnostic features with DAT to overcome the domain mismatch problem. The experiments show that the proposed method delivers voice quality comparable with other state-of-the-art supervised learning techniques.

The second proposed method addresses the mismatch problem caused by missing high-frequency signals (i.e., channel effect), which is commonly seen in radio-channel corpus. Under such scenarios, input signals are noisy, as well as lacking

high-frequency information due to the channel effect. To recover the missing information and also reduce background noises, we combine speech enhancement techniques and bandwidth extension with multi-task learning. Specifically, we propose an end-to-end time-domain framework for noise-robust bandwidth extension, that jointly optimizes mask-based speech enhancement and the bandwidth extension module with a multi-task loss function. In addition, the proposed framework also avoids decomposing signals into magnitude and phase spectra, and therefore requires no phase estimation. Experimental results show that the proposed method achieves better performance over the best baseline with fewer parameters.

The third proposed method addresses the mismatch problem caused by sensitive time-domain encoder/decoder. Time-domain speech enhancement has recently made great progress thanks to the learned filterbanks in the speech encoder/decoder as used in Conv-TasNet [1]. However, the learned filterbanks in the encoder/decoder are usually trained by fully relying on the training data, which are sensitive to unseen test data. To alleviate this problem, we propose a two-step hybrid filterbanks-based network (TSHFNet) consisting of the fully-learned filters, semi-learned filters, and non-learned filters that can improve the robustness of the speech encoder/decoder when faced with matched/unmatched testing environments. The experiments confirm that the proposed method is more robust than the best time-domain speech enhancement baseline.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Humans have a remarkable ability to focus their attention on the words of a particular speaker, even in noisy acoustic environments where there are other people speaking at the same time or where there are other noises present. Speech enhancement algorithms mimic the human ability to selectively pay attention, and mask out background noise, in order to focus on important speech content. Such algorithms have served as a pre-processing module in many real-world applications, such as automatic speech recognition (ASR), speaker identification, and hearing aids design [8, 9].

A brief history of speech enhancement systems. For several decades, various statistical approaches were proposed to mimic the human ability to selectively pay attention by masking out background noise, such as subspace algorithms [10], minimum-mean-square-error (MMSE)-based spectral amplitude estimator [11], soft-decision noise suppress filter [12], generalized gamma priors [13] and others [14–20]. Such algorithms attempt to estimate an optimal multiplicative masking, through statistical inference, in order to suppress background noises.

With the advent of deep neural networks (DNNs), learning-based models were applied to predict clean speech from distorted inputs in frequency domain, such as feed-forward networks [21–24], recurrent and long short-term memory (LSTM) networks [2, 25–27]. Some studies [1, 28] reveal that frequency-domain methods

have several limitations. First, short-time Fourier transform (STFT) is a general signal transformation which is not proved as the most optimal feature extraction for speech enhancement. Second, accurate reconstruction of the phase for enhanced speech is a difficult problem, and the erroneous estimation of the phase leads to sub-optimal speech quality [29, 30]. Some post-processing methods are proposed to alleviate this problem by predicting phase information [31–35], but the final performance is still sub-optimal.

More recently, end-to-end frameworks were proposed to avoid the phase manipulation issue by operating directly on the noisy waveform and eliminating an explicit STFT, such as fully convolutional networks (FCNs) [36–38], Wave-U-Net architectures [39–43], Conv-Tasnet [1] and others [27, 36, 37, 44, 45]. Such time-domain end-to-end approaches can optimize the whole encoder-decoder like structure as well as omit the phase estimation.

The mismatch problem. However, for learning-based speech enhancement frameworks, we usually assume that training and testing data have the same probability distribution. Real-life scenarios often fail to meet this assumption. Therefore, speech enhancement performance may degrade significantly when faced with mismatched scenarios at run-time. This research addresses the mismatch problem in speech enhancement models. Many factors cause this problem in speech enhancement, such as unseen speakers, unseen accents, *unseen noises in the test data, channel distortion and the sensitive time-domain encoder/decoder*. In this thesis, we only focus on the last three factors.

A brief review of related work. To address such mismatch problem, several deep learning methods were proposed [3, 4, 46, 47]. We briefly introduce several techniques which mainly aim to solve the problem caused by the aforementioned three factors.

1) To address the mismatch problem caused by unseen noises in the test data, one popular approach is to adapt the speech enhancement model, which is trained with the source domain data, to the test data (i.e., the target domain data). For example, the study in [46] suggests adapting the last layers of pre-trained speech enhancement generative adversarial network (SEGAN) with the dataset of new noises to reduce the mismatch between different noises.

However, this technique requires clean-noisy parallel speech data that are not always available in practice.

2) To address the mismatch problem caused by channel distortion, prior work [47] combines speech enhancement and bandwidth extension (named UEE) in a joint training neural network. The UEE is first trained separately at the pre-training stage, and then fine-tuned with a single mean square error (MSE) loss function. Overall, the UEE approach is implemented with the pre-training scheme, and it also faces phase estimation difficulties just like other frequency-domain techniques.

3) To address the mismatch problem caused by sensitive time-domain encoder/decoder, prior work [3] attempts to directly combine the frequency-domain network and time-domain network to obtain balanced results, by averaging the two outputs respectively. Likewise, work [4] combines the frequency-domain network and time-domain network in a sequence order to take advantage of the two-domain features. The performance of such methods remains sub-optimal.

These prior studies are the source inspiration of the proposed novel approaches in this thesis.

1.2 Contributions

The thesis provides three novel speech enhancement methods when faced with the mismatch problem. The first method addresses the mismatch problem caused by unseen noises in the test data, the second method addresses the mismatch problem caused by channel distortion and the third method addresses the mismatch problem caused by sensitive time-domain encoder/decoder. Three mismatched scenarios are different, but they can be grouped into general training-testing mismatch problem.

1.2.1 Speech enhancement with adversarial training

The performance of deep learning approaches to speech enhancement degrades significantly in the face of noises mismatch between training (source-domain) and

testing (target-domain). We first focus on the mismatch problem caused by unseen noises in the test data. To alleviate this problem, we apply adversarial training strategy to speech enhancement frameworks in two cases: 1) with target-domain data and 2) without target-domain data. For case 1), when the noisy target-domain data is available, we propose an unsupervised domain transfer approach by adapting the enhancement network without the need for clean-noisy parallel speech data in the target domain. In our scenarios, the training data for the source domain consists of clean-noisy speech pairs, but those in the target domain only consist of noisy speech. For case 2), when the noisy target-domain data is not available, we propose to learn noise-agnostic feature representations through disentanglement learning (one specific application of adversarial training strategy), which removes the unspecified noise factor, while keeping the specified factors of variation associated with the clean speech. Experimental results show that the proposed method achieves better performance over the best baseline with target-domain data and without target-domain data. The proposed methods will be introduced in Chapter 3.

The proposed methods were documented and published in:

1) Hou, N., Xu, C., Chng, E.S. and Li, H., 2019, November. Domain adversarial training for speech enhancement. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 667-672). IEEE.

2) Hou, N., Xu, C., Chng, E.S. and Li, H., 2021, June. Learning Disentangled Feature Representations for Speech Enhancement Via Adversarial Training. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 666-670). IEEE.

1.2.2 Speech enhancement with multi-task learning

We next explore the mismatch problem caused by missing high-frequency information under radio-channel condition (channel distortion). Prior studies mostly perform speech enhancement under the assumption that the signals are only corrupted by noise but are not lost. The use of such speech enhancement techniques is greatly limited in practice under channel condition where high-frequency signals in noisy

speech are usually missing due to distortion. To alleviate this problem, we propose an end-to-end time-domain framework for noise-robust bandwidth extension, which jointly optimizes the mask-based speech enhancement and the bandwidth extension modules with a multi-task learning (MTL-MBE). As a time-domain technique, the proposed method inherently avoids phase estimation issues. Experimental results show that the proposed method significantly improves the performances of speech quality and intelligibility over the best baseline (UEE). The proposed method will be provided in Chapter 4.

The proposed method was documented and published in:

1) Hou, N., Xu, C., Zhou, J.T., Chng, E.S. and Li, H., 2020. Multi-Task Learning for End-to-End Noise-Robust Bandwidth Extension. In INTERSPEECH (pp. 4069-4073).

1.2.3 Speech enhancement with hybrid filterbanks

In this study, we focus on the mismatch problem caused by the sensitive time-domain encoder/decoder. Recently, time-domain speech enhancement is widely used to alleviate the phase reconstruction problem thanks to the learned filterbanks as used in Conv-TasNet [1]. However, this approach is usually trained by fully relying on the training data, which is sensitive to varying test data [16, 48, 49]. Analyses of recent studies [50–52] suggest that the cause of this problem is two-folds. First, convolutional filters in the time-domain speech encoder fully rely on training data. They cannot always yield a good decomposition of the input speech as compared with the fixed STFT algorithm in practice. Second, the end-to-end training strategy cannot guarantee that the speech encoder and decoder are well trained in this process, which might also cause the sensitivity of time-domain speech enhancement frameworks. To alleviate this problem, we design a two-step hybrid filterbank including the fully-learned filters, semi-learned filters and the non-learned filters for time-domain speech enhancement (TSHFNet) to improve robustness when faced with the varying testing environments. Experimental results show that the proposed TSHFNet performs better than the best baseline Conv-TasNet on the DNS corpus. The proposed method will be provided in Chapter 5.

The proposed method was documented and published in:

1) Hou, N., Chng, E.S. and Li, H., 2021. Hybrid Filterbanks with Two-Step Training for Time Domain Speech Enhancement. Submitted to ACM Transactions on Audio Speech and Language Processing 2021.

1.3 Thesis organization

We organize the thesis into six chapters as follows:

Chapter 1 introduces the motivation, the contributions, and the brief summary of the thesis.

Chapter 2 first introduces background for speech enhancement. Then, we review state-of-the-art speech enhancement techniques and the existing problem of mismatched probability distributions between training data and test data.

In Chapter 3, we apply the adversarial training strategy to speech enhancement techniques in two cases: with target-domain data and without target-domain data, to alleviate the mismatch problem caused by the unseen noises in the test data. The detailed description, experimental setup and evaluation are provided in this chapter.

In Chapter 4, we introduce the multi-task learning scheme to alleviate the mismatch problem caused by channel distortion. We start with a description of the limitation of prior work. We then present our proposed solution. Finally, we clarify the experimental setup and analysis the experimental results.

In Chapter 5, we present a novel hybrid filterbanks design for the mismatch problem caused by the sensitive time-domain encoder/decoder. We first describe the limitations of time-domain approaches and then introduce the proposed method. Finally, we present the experimental setup and discuss the experimental results.

Chapter 6 concludes contributions of this thesis, and provides future directions.

Chapter 2

Literature Review of Speech Enhancement

This chapter begins with a definition of signals and an introduction to speech enhancement in Section 2.1. Since speech enhancement techniques have been explored for decades, we review some of the previous representative studies in Section 2.2 and Section 2.3. Specifically, Section 2.2 summarizes traditional speech enhancement technologies prior to the deep learning era. Section 2.3 reviews deep-learning-based speech enhancement technologies. In addition, limitations of the current speech enhancement work are discussed in Section 2.4.

2.1 Background

The human brain has a strong ability to focus on speech content of interest in noisy acoustic situations, such as when multiple speakers, competing to be heard, contribute to the ambient noise. Speech enhancement algorithms attempt to duplicate the selective attention of the human brain in masking out background noises and focus on important speech content. Such algorithms are usually applied as a pre-processing module in many real-world applications, such as automatic speech recognition (ASR), speaker identification, and hearing aids design [8, 9].

2.1.1 Definition of speech enhancement

Speech signals recorded with distant microphones (i.e. microphones placed some distance from the source) are inevitably corrupted by various noises, which severely degrade the perceptual quality of the recorded speech signals. If we have a single distant microphone as the recording device, the received discrete signal $y(n)$ will be expressed as,

$$y(n) = s(n) + d(n) \tag{2.1}$$

where n is a time position of each sample. $s(n)$ denotes the clean speech and $d(n)$ is the background noise.

Speech enhancement aims to reduce the additive noises $d(n)$ in the noisy signals $y(n)$ in order to improve speech quality and intelligibility. To estimate the clean signals $s(n)$, most speech enhancement methods first transform time-domain noisy signals into their magnitude and phase spectra with a short time Fourier transform (STFT) to facilitate frequency-domain processing.

By using a STFT, the time-domain noisy signals $y(n)$ is transformed as,

$$Y(t, f) = \sum_{n=0}^{N-1} y(n + tL)w(n)e^{(\frac{-j2\pi nf}{N})} \tag{2.2}$$

where a time frame $t \in [0, T-1]$ is obtained by shifting an amount of L samples, and a frequency bin $f \in [0, T-1]$ is associated with a frequency of $\frac{f}{N}f_s$ Hz at a sampling rate of f_s Hz. $w(n)$ denotes the Hamming or Hanning window of length N.

With the time-domain signal model as Equation 2.1, the frequency-domain noisy signals can be decomposed as,

$$Y(t, f) = S(t, f) + D(t, f) \tag{2.3}$$

where $S_c(t, f)$ and $D(t, f)$ are the complex spectra of the corresponding time-domain clean signals and background noises, respectively.

In a typical setup, speech enhancement estimates the complex spectrum $\hat{S}$, which is close to the original clean spectrum S. To obtain the complex spectrum $\hat{S}$, the

magnitude $|\hat{S}|$ and phase $\measuredangle\hat{S}$ spectra need to be estimated. Since phase spectrograms show little temporal and spectral regularities, it is a challenging problem to transform the phase spectrum into a better one. Most speech enhancement methods use the phase spectrogram of the noisy signals as the phase for the estimated spectrogram ($\measuredangle\hat{S} = \measuredangle Y$). Therefore, frequency-domain speech enhancement methods typically focus on how to improve the quality of the estimated magnitude spectrum $|\hat{S}|$.

The goal of the speech enhancement is to make the estimated magnitude $|\hat{S}|$ as close as possible to the clean magnitude $|S|$. When the magnitude spectrum $|\hat{S}|$ is estimated, the complex spectrogram $\hat{S}$ could be obtained by combining the magnitude $|\hat{S}|$ and the phase $\measuredangle Y$ of the noisy signals. Therefore, the estimated time-domain signal $\hat{s}$ could be reconstructed by an overlap and add algorithm,

$$\hat{s}(n) = \sum_{t=0}^{T-1} v(n - tL)\hat{s}_t(n - tL) \tag{2.4}$$

together with an inverse STFT (iSTFT),

$$\hat{s}_t(n) = \frac{1}{N} \sum \hat{S}(t, f)e^{(\frac{j2\pi n}{N})} \tag{2.5}$$

where $v(n)$ is a window similar to $w(n)$ in Equation 2.2

2.1.2 Corpus

Recent works are mostly conducted using publicly available corpus for easy comparisons, such as noisy speech database for training speech enhancement algorithms and TTS models (Valentini database), English multi-speaker corpus for CSTR Voice Cloning Toolkit (VCTK database), and CHiME speech Separation and recognition challenge (CHiME series database). This thesis also conducts experiments using these well-known speech enhancement corpora in order to have a fair comparison with state-of-the-art speech enhancement methods.

2.1.2.1 Valentini database

The valentini database[1] artificially adds the recorded noises to clean speech as noisy data, and forms the noisy/clean pairs for speech enhancement training. Specifically, clean data is selected from the Voice Bank corpus [53], including two subsets. One subset includes 28 speakers: 14 male and 14 female from the same accent region (England) and the other subset has 56 speakers: 28 male and 28 female from different accent regions (Scotland and United States). There are around 400 sentences available from each speaker. All data is sampled at 48 kHz.

Meanwhile, the Valentini database utilizes ten different types of noise: two artificially generated (speech-shaped noise and babble) and eight real noise recordings from the Demand database [54]. The signal-to-noise (SNR) values used for the simulation were: 15 dB, 10 dB, 5 dB and 0 dB. We therefore had 40 different noisy conditions (ten noises $\times$ four SNRs), which meant that for each speaker there were around ten different sentences in each condition.

For the database of noisy recordings used for testing, the Valentini database selected two unseen speakers from England, a male and a female of the same corpus, and five unseen noises from the Demand database. The new noises includes a domestic noise (living room), an office noise (office space), one transport (bus) noise and two street noises (open area cafeteria and a public square). The SNR values were set at: 17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB. This created 20 different noisy conditions (five noises $\times$ four SNRs), which meant that for each speaker there were around 20 different sentences in each condition. The noise was added following the same procedure described previously.

2.1.2.2 VCTK database

The VCTK database[2] was created to simulate a larger clean/noisy parallel corpus for the speech enhancement training. This CSTR VCTK database includes speech data uttered by 110 English speakers with various accents[3]. Each speaker reads out about 400 sentences, selected from a newspaper, the rainbow passage and an elicitation paragraph used for the speech accent archive. The newspaper texts were

[1] Available at https://datashare.ed.ac.uk/handle/10283/1942
[2] Available at https://datashare.ed.ac.uk/handle/10283/3443
[3] Available at `http://www.ualberta.ca/~aacl2009/PDFs/WeinbergerKunath2009AACL.pdf`

taken from the Herald Glasgow, with permission from the Herald Times Group. Each speaker has a different set of newspaper texts selected based on a greedy algorithm that increases the contextual and phonetic coverage. The rainbow passage[4] and elicitation paragraph[5] are the same for all speakers.

All speech data was recorded using an identical recording setup: an omni-directional microphone (DPA 4035) and a small diaphragm condenser microphone with a very wide bandwidth (Sennheiser MKH 800), 96kHz sampling frequency at 24 bits. All speech data were recorded in a hemi-anechoic chamber of the University of Edinburgh. All recordings were converted into 16 bits, were downsampled to 48 kHz, and were manually end-pointed. This corpus was originally meant for text-to-speech synthesis systems, especially for speaker-adaptive speech synthesis that used average voice models trained on multiple speakers and speaker adaptation technologies. Recently, this corpus was also widely used for speech enhancement training.

2.1.2.3 CHiME series database

The $1^{st}-6^{th}$ CHiME series database considered the problem of speech enhancement and conversational speech recognition in everyday home environments. Here, we take the 4^{th} CHiME database as an example because it is utilized in Chapter 3.

The CHiME-4 database consists of two subsets: a real dataset and a simulated dataset. The real dataset utilized a multi-microphone tablet device to record data in everyday noisy environments, which represents a significant step forward in terms of realism compared with the CHiME-1 and CHiME-2 challenges. Specifically, the clean utterances of the CHiME-4 database are provided in continuous audio with ground truth VAD annotations. The original live recordings were made by 12 American speakers (6 male and 6 female). For each speaker, recordings were made first in a sound proof booth and then in each of the four target noisy environments: cafes, street junctions, on public transport and at pedestrian areas. 100 sentences were read at each location. It was stressed that each sentence had to be read correctly and without interruption. Speakers were allowed as many attempts as necessary to read each sentence. They were asked to use the tablet in whatever

[4]Available at `http://web.ku.edu/~idea/readings/rainbow.htm`
[5]Available at `http://accent.gmu.edu`

way felt natural and comfortable. They were encouraged to adjust their reading position after every 10 utterances, e.g. holding the tablet (most typical), resting it on their lap, laying it on a table, etc. Simulated data was created by artificially mixing clean speech data (WSJ0 SI-84) with background noises (live recordings).

All the data in the CHiME-4 database was divided into a training set, a development test and a test sets. Each set has recordings from different speakers and different noisy environments. For example, all data sets have recordings from the noisy cafe environment but different specific cafes are used in each set. The audio data is provided as 16 bit stereo WAV files sampled at 16 kHz. The details of the three sets are listed as follows:

- Training set: 1600 (real) + 7138 (simulated) = 8738 noisy utterances from a total of 4 speakers in the real data, and 83 speakers from the WSJ0 SI-84 training set in the 4 noisy environments. The transcriptions are also corresponded to the speech utterances in the WSJ0 SI-84 training set, but the real speech utterances contain no verbal punctuations (e.g., "period" and "hyphen" in the original WSJ0 SI-84). All of the reading errors in these transcriptions are corrected appropriately.
- Development set: 410 (real) × 4 (environments) + 410 (simulated) × 4 (environments) = 3280 utterances from 4 new speakers, who are different from those in the training set. The utterances are based on the "no verbal punctuation" (NVP) part of the WSJ0 speaker-independent 5k vocabulary development set.
- Test set: 330 (real) × 4 (environments) + 330 (simulated) × 4 (environments) = 2640 utterances from 4 new speakers, who are different from those in the training set and the development set. Similar to the development set, the utterances are based on the "no verbal punctuation" (NVP) part of the WSJ0 speaker-independent 5k vocabulary evaluation set.

2.1.3 Evaluation metrics

Speech enhancement systems can be evaluated by both objective evaluation and subjective evaluation. Objective evaluation is used to measure the distortion and

quality between the enhanced and clean signals. Subjective evaluation examines the perceptual quality and intelligibility of the enhanced signals.

2.1.3.1 Objective evaluation

Perceptual Evaluation of Speech Quality (PESQ) [50, 51] is recommended as the ITU-T P.862 standard to automatically assess speech quality instead of the subjective Mean Opinion Score (MOS). The key to using PESQ is to predict the MOS value that would result if real people were evaluating the recorded speech clips. Compared with MOS, PESQ has the benefit of avoiding the involvement of listening subjects in the evaluation. Instead of using the signal-to-distortion ratio without distinguishing the type of distortion, PESQ incorporates a perceptual model to distinguish between audible distortion (i.e. a noise added to the spectrum) and inaudible distortion (i.e. a spectral component omitted or heavily attenuated) with different weights. Compared to additive components, the omitted or attenuated components may not be easily perceived because of masking effects. The range of the PESQ score is –0.5 to 4.5. The higher the score, better the performance of the speech enhancement systems.

Likewise, the other three objective metrics that approximate MOS [55] are: CSIG, CBAK and COVL. These metrics are designed for signal distortion evaluation, noise distortion evaluation, and overall quality evaluation, respectively. Short-time objective intelligibility (STOI) [56] reflects the improvement of speech intelligibility.

Signal-to-distortion ratio (SDR) is also conducted for the purpose of measuring the speech quality, which calculates the energy ratios expressed in decibels between the estimated wanted signal and the distortion, the interferences and the artifacts, respectively. The ratios are calculated by first decomposing the estimated signal $\hat{s}$ as

$$\hat{s} = s_t + e_i + e_n + e_a \tag{2.6}$$

where $s_t = f(s)$ is the modified source signal s with an allowed distortion $f(\cdot)$. e_i, e_n, and e_a are the interference, sensor noise, and burbling artifacts error terms, respectively.

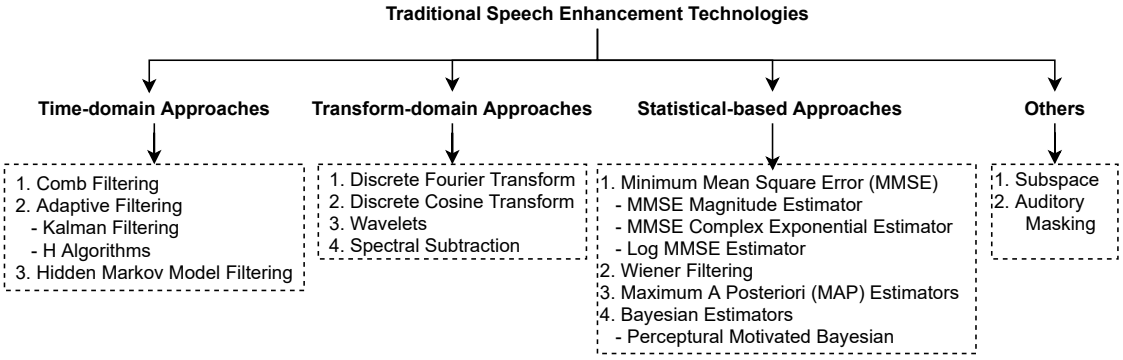


FIGURE 2.1: The classification of the traditional speech enhancement technologies into four categories: time-domain approaches, transform-domain approaches, statistical-based approaches and others.

SDR is then calculated as,

$$SDR = 1 - log_{10}\frac{||s_t||^2}{||e_i + e_n + e_a||^2} \tag{2.7}$$

2.1.3.2 Subjective evaluation

The A/B preference test is a subjective evaluation comparing speech quality between two systems. For each A/B listening pair, one sample is from the proposed system while the other one is from the system used as the baseline for comparison. These two samples are randomly presented to listeners. The listeners don't know which sample belongs to which system. Each listener is asked to listen to both samples and chooses the better sample in terms of speech quality. The higher the preference rate for the samples of a speech enhancement systems, the better the performance of that system.

2.2 Traditional speech enhancement technologies

For several decades, common speech enhancement technologies prior to the deep learning era used various signal processing tools to reduce background noises. Based on the type of processing used, we classify these traditional speech enhancement technologies into four categories: time-domain approaches, transform-domain approaches, statistical-based approaches and others, as shown in Table 2.1.

Time-domain approaches. Time-domain speech enhancement approaches are directly performed on speech waveforms via the application of various designed filters, such as comb filtering [57], adaptive filtering [58, 59] and hidden Markov model filtering [60]. In statistics and control theory, filtering methods are algorithms that uses a series of measurements observed over time, including statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone. This is done by estimating a joint probability distribution over the variables for each time frame. For example, the Kalman filtering [58] is a two-phase process. For the prediction phase, the Kalman filter produces estimates of the current state variables, along with their uncertainties. Once the outcome of the next measurement is observed, these estimates are updated using a weighted average, with more weight being given to estimates with greater certainty. The algorithm is recursive. It can operate in real time, using only the present input measurements and the state calculated previously and its uncertainty matrix. No additional past information is required.

Transform-domain approaches. Transform-domain speech enhancement approaches first revert time-domain speech signals into transform-domain representations via discrete Fourier transform [61] and discrete cosine transform [62]. In the frequency domain approaches, a short-time Fourier transform (STFT) [63] is typically applied to a windowed time-domain noisy speech signal to extract stable acoustic features for subsequent analysis. After the transformation, various spectral estimation methods are applied to reduce unnecessary noise. AS spectral subtraction [64] is a well-known method to reduce background noise in noisy speech signals, we take it as an example. Assuming the given noise signals are additive and are uncorrelated, the spectral subtraction method first estimates the noise spectrum in the noisy speech spectrum and then subtracts the estimated noise spectrum from the noisy speech spectrum while keeping the phase undisturbed. Hence an estimate of background noise obtained from the region of noisy speech, is subtracted from the noisy speech to obtain the enhanced speech signal.

Statistical-based approaches. Statistical speech enhancement approaches utilize asymptotic statistical properties of the Fourier expansion coefficients to derive a spectral amplitude estimator, such as the family of minimum mean square error algorithms [11, 14, 65], Wiener filtering [66, 67], maximum a posteriori [68]

and Bayesian estimator [69]. For example, the MMSE method [11] estimates the modulation magnitude spectrum of clean speech from noisy observations and the estimator minimises the mean-square error between the modulation magnitude spectra of clean and estimated speech. In the Wiener filtering method [66], a spectral gain function of SNR measures is computed and applied to the noisy speech to get an estimate of clean speech.

Others. Some speech technologies do not belong to the above three categories, such as signal subspace methods [70] and the auditory masking method [71, 72]. Signal subspace methods essentially represent the application of a principal component analysis approach to ensembles of observed time-domain signals obtained by sampling. Auditory masking is utilized when the perception of one sound is affected by the presence of another sound. Masking can be simultaneous or non-simultaneous, where a masked threshold is setup to control the sound to be heard or unwanted.

In this subsection, we will introduce two representative traditional approaches: the subspace algorithms and MMSE-based spectral amplitude estimator.

2.2.1 Spectral subtraction approach

We now introduce the spectral subtraction (SS) approach [64], as shown in Figure 2.2. Consider a noisy signal which consists of the clean speech degraded by statistically independent additive noise as,

$$y(n) = s(n) + d(n) \tag{2.8}$$

where $y(n), s(n)$ and $d(n)$ are the sampled noisy speech, clean speech, and additive noise, respectively. It is assumed that additive noise is zero mean and uncorrelated with the clean speech. As the speech signal is non-stationary and time variant, the noisy speech signal is often processed on a frame-by-frame.Their representation in the short-time Fourier transform (STFT) domain is given by,

$$Y(w, k) = S(w, k) + D(w, k) \tag{2.9}$$

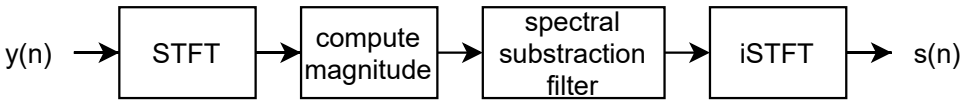


FIGURE 2.2: The block diagram of the spectral subtraction approach.

where k is a frame number. As the speech signal is segmented into frames, for simplicity, we drop k. Since the speech is assumed to be uncorrelated with the background noise, the short-term power spectrum of $y(n)$ has no cross-terms. Hence,

$$|Y(w)|^2 = |S(w)|^2 + |D(w)|^2 \tag{2.10}$$

The speech can be estimated by subtracting a noise estimate from the received signal.

$$|\hat{S}(w)|^2 = |Y(w)|^2 - |\hat{D}(w)|^2 \tag{2.11}$$

The estimation of the noise spectrum $|\hat{D}(w)|^2$ is obtained by averaging recent speech pauses frames:

$$|\hat{D}(w)|^2 = \frac{1}{M}\sum_{j=0}^{M-1}|Y_{SP_j}(w)|^2 \tag{2.12}$$

where M is the number of consecutive frames of speech pauses (SP). If the background noise is stationary, Equation 2.12 converges to the optimal noise power spectrum estimate as a longer average is taken.

The spectral subtraction can also be looked at as a filter, by manipulating Equation 2.11 such that it can be expressed as the product of the noisy speech spectrum and the spectral subtraction filter (SSF) as:

$$\begin{aligned} |\hat{S}(w)|^2 &= (1 - \frac{|\hat{D}(w)|^2}{|Y(w)|^2})|Y(w)|^2 \\ &= H^2(w)|Y(w)|^2 \end{aligned} \tag{2.13}$$

where $H(w)$ is the gain function and known as SSF. $H(w)$ is a zero phase filter, with its magnitude response in the range of $0 \leq H(w) \leq 1$.

$$H(w) = max(0, 1 - \frac{|\hat{D}(w)|^2}{|Y(w)|^2})^{\frac{1}{2}} \tag{2.14}$$

To reconstruct the resulting signal, the phase estimate of the speech is also needed. A common phase estimation method is to adopt the phase of the noisy signal as the phase of the estimated clean speech signal, based on the notion that short-term phase is relatively unimportant to human ears. Then, the speech signal in a frame is estimated as,

$$\hat{S}(w) = |\hat{S}(w)|e^{j\leq Y(w)} = H(w)Y(w) \tag{2.15}$$

The estimated speech waveform is recovered in the time domain by inverse Fourier transforming $\hat{S}(w)$ using an overlap and add approach.

The spectral subtraction method, although it reduces noises significantly, it has some severe drawbacks. From Equation 2.11, it is clear that the effectiveness of spectral subtraction is heavily dependent on accurate noise estimation, which is a difficult task to achieve in most conditions. When the noise estimate is less than perfect, two major problems occur, namely remnant noise with musical structure and speech distortion.

Spectral subtractive type algorithms are the family of different variants of the spectral subtraction method, such as spectral over-subtraction, multi-band spectral subtraction, Wiener filtering, iterative spectral subtraction, and spectral subtraction based on perceptual properties. Thus, the primary aim of spectral subtractive type algorithms is to estimate the short-time spectral magnitude of speech by subtracting estimated noise from the noisy speech spectrum, or by multiplying the noisy spectrum with gain functions and to combine it with the phase of the noisy speech.

2.2.2 Minimum mean square error based approach

Spectral subtraction is one of the earliest and most extensively studied methods for speech enhancement. This simple method enhances speech by subtracting a spectral estimate of noise from the noisy speech spectrum in either the magnitude or energy domain. Though this method is effective at reducing noise, it suffers from the problem of musical noise distortion, which is very annoying to listeners. To overcome this problem, work [11] proposed the MMSE short-time spectral amplitude estimator.

In the MME method, the modulation magnitude spectrum of clean speech is estimated from noisy observations. The proposed estimator minimises the mean-square error between the modulation magnitude spectra of clean and estimated speech

$$\epsilon = E[(|S_l(k,m)| - |\hat{S}_l(k,m)|^2)^2] \tag{2.16}$$

where $E[\cdot]$ denotes the expectation operator. The assumption of the MME based approach is that:

- Speech and noise are additive in the time domain.
- Their individual short-time spectral components are statistically independent, identically distributed, zero-mean Gaussian random variables.

The assumptions are formulated as follows:

$$|Y_l(k)| = |S_l(k)| + |D_l(k)| \tag{2.17}$$

where $S_l(k,m)$ and $D_l(k,m)$ are independent individual short-time modulation spectral components in Gaussian distribution.

- The reasoning for the first assumption is that at high SNRs the phase spectrum remains largely unchanged by additive noise distortion [73].
- For the second assumption, it follows the configurations with the prior work [11], where the central limit theorem is used to justify the statistical independence of spectral components of the Fourier transform. This assumption is valid only in the asymptotic sense for the STFT, that is, when the frame duration is large. However, work [11] used an acoustic frame duration of 32 ms in their formulation to get good results. Therefore, the MMSE approach should also make the modulation frame duration to be as large as possible, however it must not be so large as to be adversely affected by the non-stationarity of the magnitude spectral sequence.

With the above assumptions in mind, the modulation magnitude spectrum of clean speech can be estimated from the noisy modulation spectrum under the MMSE

criterion [11] as

$$\begin{aligned} |\hat{S}_l(k,m)| &= E[|S_l(k,m)||Y_l(k,m)|] \\ &= G_l(k,m)|Y_l(k,m)| \end{aligned} \tag{2.18}$$

where $G_l(k,m)$ is the MMSE-MME spectral gain function given by,

$$\begin{aligned} G_l(k,m) &= \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_l(k,m)}}{r_l(k,m)}A[v_l(k,m)] \\ v_l(k,m) &= \frac{\xi_l(k,m)}{1+\xi_l(k,m)}\gamma_l(k,m) \\ A[\theta] &= EXP(-\frac{\theta}{2})[(1+\theta)i_0(\frac{\theta}{2}+\theta I_1(\frac{\theta}{2}))] \\ \xi_l(k,m) &= \frac{E[|S_l(k,m)|^2]}{D[|S_l(k,m)|^2]} \\ \gamma_l(k,m) &= \frac{|Y_l(k,m)|^2}{D[|S_l(k,m)|^2]} \end{aligned} \tag{2.19}$$

where $I_0(\,\cdot\,)$ and $I_1(\,\cdot\,)$ denote the modified Bessel functions of zero and first order, respectively. In the above equations $\xi_l(k,m)$ and $\gamma_l(k,m)$ are interpreted as the a priori SNR, and the a posteriori SNR [12], respectively. Since in practice only noisy speech is observable, the $\xi_l(k,m)$ and $\gamma_l(k,m)$ parameters have to be estimated. For this task, the decision-directed approach was applied to the short-time spectral modulation domain [11].

This non-parametric based approach is formulated to estimate the clean speech spectrum by subtracting an estimate of the noise spectrum from the observation spectrum. The main drawback of this approach is the appearance of unnatural sounding artifacts, known as musical noise, which is annoying and unpleasant to the listeners.

2.3 Deep learning based speech enhancement technologies

With the advent of deep neural networks (DNNs), learning-based models were applied to process speech in order to derive clean speech from distorted inputs. Such

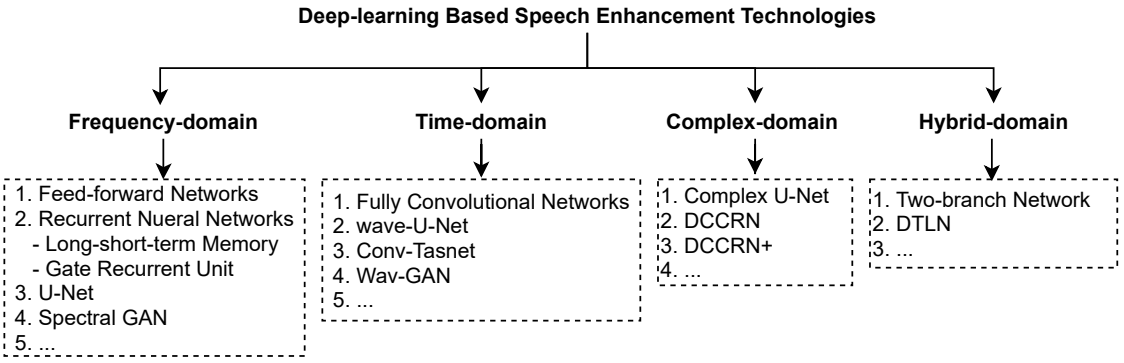


FIGURE 2.3: The classification of the deep-learning-based speech enhancement technologies into four categories: frequency-domain approaches, time-domain approaches, complex-domain approaches and hybrid-domain approaches.

models can be grouped into four categories as shown in Figure2.3: frequency-domain approaches, time-domain approaches, complex-domain approaches and hybrid-domain approaches.

Frequency-domain approaches. Deep learning based technologies were firstly explored in the frequency domain, such as feed-forward networks [21–24], recurrent and long short-term memory (LSTM) networks [2, 25–27], U-Net [74] and spectral generative adversarial network (GAN) [75], as shown in Figure 5.1 (a). Some studies [1, 28] reveal that frequency-domain methods have several limitations. First, short-time Fourier transform (STFT) is a general signal transformation which is not proved as the most optimal feature extraction for speech enhancement. Second, accurate reconstruction of the phase for enhanced speech is a difficult problem, and the erroneous estimation of the phase leads to sub-optimal speech quality [29, 30]. Some post-processing methods are proposed to alleviate this problem by predicting phase information [31–35], but the final performance is still sub-optimal.

Time-domain approaches. More recent end-to-end frameworks were proposed to avoid the phase manipulation issue by operating directly on the noisy waveform and eliminating an explicit STFT, such as fully convolutional networks (FCNs) [36–38], wave-U-Net architectures [39–43], Conv-Tasnet [1] and others [27, 36, 37, 44, 45], as shown in Figure 5.1 (b). Such time-domain end-to-end approaches can optimize the whole encoder-decoder-like structure as well as omit the phase estimation. However, as the time-domain encoder and decoder are fully learned from the training data, they are sensitive to varying testing environments [16, 48, 49].

Recent studies also [50–52] observed that the time-domain speech separation/enhancement frameworks might not always yield good performances when faced with varying testing scenarios. The problem is in two-folds. First, the convolutional filters in the time-domain speech encoder fully rely on training data. They cannot always yield a good decomposition of the input speech compared with the fixed STFT algorithm when faced with various test data. Second, the end-to-end training strategy cannot guarantee that the speech encoder and decoder are well trained in this process, which might also cause the sensitivity of time-domain speech separation/enhancement frameworks.

Complex-domain approaches. To alleviate this problem, prior works [7, 76] attempted to design a deep complex convolution recurrent network (DCCRN and DCCRN+) to optimize a SI-SNR loss, which take complex values of acoustic features as inputs. The networks effectively combines the advantages of frequency-domain acoustic features with the imaginary value of features to predict phase information. Likewise, another work [77] extends the U-Net architecture and also utilizes complex values of acoustic features as inputs, named complex U-Net.

Hybrid-domain approaches. Another way is to [3] directly combine the frequency-domain network and time-domain network as two branches to obtain the balanced results by averaging the two outputs respectively from the two branches, as shown in Figure 5.1 (c). Likewise, work [4] combines the frequency-domain network and time-domain network in a sequence to take advantages of two domains features, as shown in Figure 5.1 (d). However, such methods have large parameters and their performance remains sub-optimal.

In this section, we will take four examples to introduce the four types of deep-learning-based speech enhancement models including: the frequency-domain approach, time-domain approach, complex-domain approach, and multi-domain approach.

2.3.1 Frequency-domain approach

We select the bi-directional long-short-term-memory (BLSTM) mask method [2] as the representative frequency-domain approach in this section. BLSTM is chosen here as it can learn future information of the audio sequences. It is a DNN

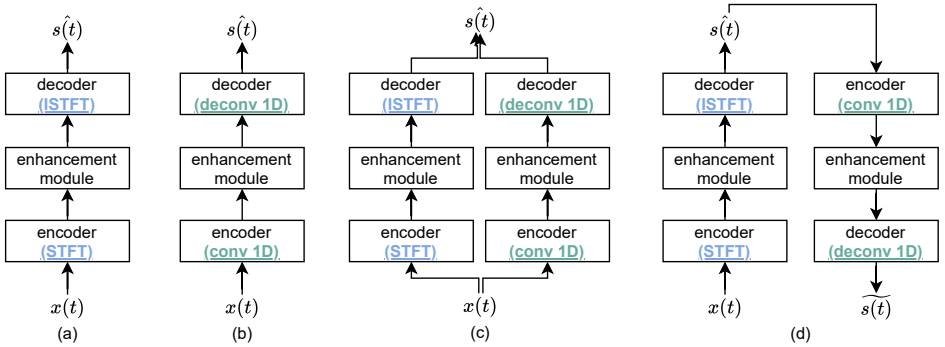


FIGURE 2.4: The block diagram of (a) frequency-domain speech enhancement network [2], (b) time-domain speech enhancement network [1], (c) two-branch multi-domain speech enhancement network [3], and (d) dual-path speech enhancement network (DTLN) [4]. $x(t)$ is the noisy speech and $\hat{s(t)}$ and $\widetilde{s(t)}$ denotes the predicted clean speech.

approach operating on spectral features with a mean square error cost function on a magnitude-spectrum. Enhancement is performed by predicting a mask to suppress unwanted signals in the spectrum domain. The final enhanced signal is reconstructed using the processed spectrum with noisy phase.

Figure 2.5 shows the overview of the BLSTM-mask approach. The input to the system is shown at the bottom of the figure. They applied 512-point STFT and half shifted size to extract features. Therefore, the magnitude spectrum features have a vertical axis of 256 (dimension), and a horizontal axis of speech frames, in the frequency domain. The system operates on utterance level using blocks of 68 speech frames (each of about 1 second). The frames are fed into 3 layers of the BLSTM module before the final linear layer which produces the same-sized 68 frames $\times$ 256 mask. This mask is combined with the noisy spectrum to produce the estimated clean spectrum. During training, the linear layer and LSTM layers are optimized to reduce mean square error between the clean and noisy spectrum.

Mish [5] is utilized as the activation function for the linear layer, which is defined as follows:

$$Mish(x) = x \times tanh(ln(1 + e^x)) \tag{2.20}$$

The curve of the Mish activation function is shown in Figure 2.6

Compared with Mish activation function, ReLU [6] has an order of continuity of zero, i.e it is not continuously differential as shown in Figure 2.7, which may cause some problems for gradient-based optimization.

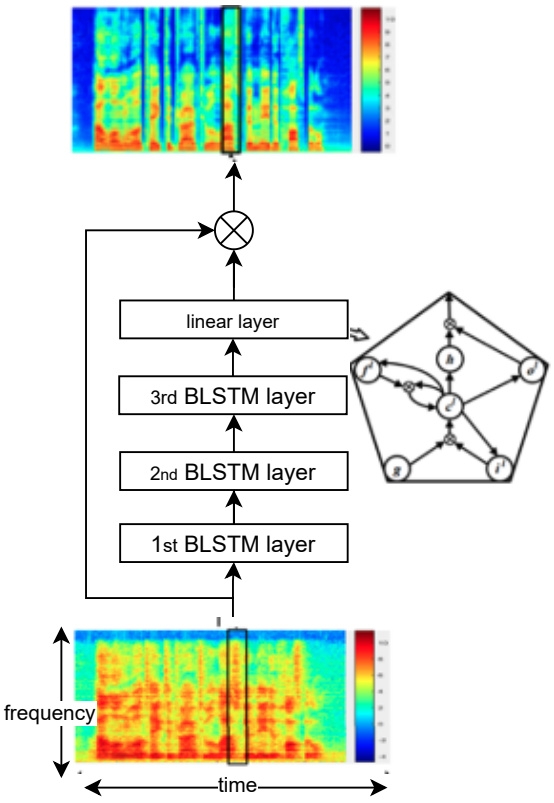


FIGURE 2.5: The block diagram of a BLSTM-mask based speech enhancement network [2]. ⊗ denotes the element-wise multiplication.

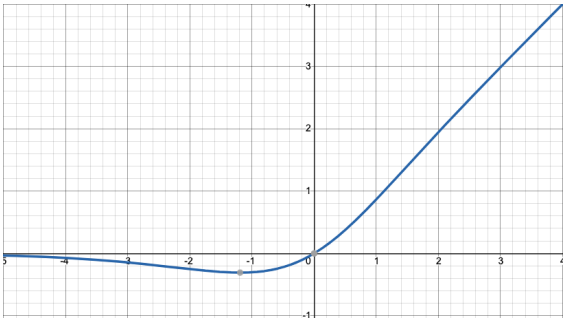


FIGURE 2.6: The illustration of the Mish activation function [5]. X-axis denotes the input values of Mish function and Y-axis is the output results of the activation function.

2.3.2 Time-domain approach

We will now introduce Conv-TasNet [1], which is proposed to solve the problem of reusing noisy inputs when reconstructing signals from spectra in frequency-domain methods. We select the Conv-TasNet [1] method as the representative time-domain approach because its good performances in speech enhancement task. The convolutional encoder and decoder were proposed to replace typical STFT and

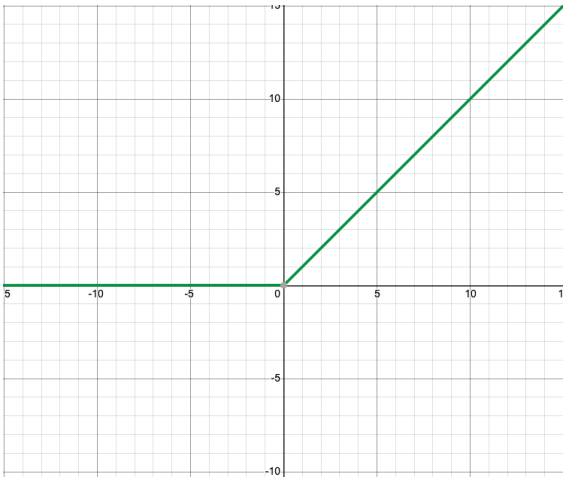


FIGURE 2.7: The illustration of the ReLU activation function [6]. X-axis denotes the input values of ReLU function and Y-axis is the output results of the activation function.

iSTFT operations in frequency-domain methods to extract spectrum-like features. The extracted spectrum-like features were then separated by the enhancement module (i.e., mask) into individual speech and recovered to signals as the decoder. The Conv-TasNet operates directly on raw waveforms and is optimized by the SI-SNR loss function. This approach shows good performance in both speech enhancement and separation tasks.

Figure 2.8 shows the structure of the Conv-TasNet. Specifically, the input noisy signal $x(t)$ is encoded to a representation A by the speech encoder. Such encoder consists of a 1-D convolutional layer, which has 512 filters with 16 samples filter size and 8 samples filter stride, followed by a rectified linear unit (ReLU) active function. Then representation A is fed into the mask estimation module, which consists of several temporal convolutional network (TCN) as illustrated in the right-hand portion of Figure 2.8.

The encoder representation is first performed zero-mean normalization on the channel dimension, and scaled by the learnable bias and gain variables. Then, such representations are adjusted the number of channels by a 1×1 convolutional layer with 128 filters. To learn the long-range temporal information of the speech with few parameters, dilated convolutional layers are stacked in several TCNs with exponentially increased dilation factor. In each TCN block, two 1×1 CNNs and one dilated convolutional layer are utilized with a parametric rectified linear unit (PReLU) activation function and normalization operation. The first 1×1 CNN

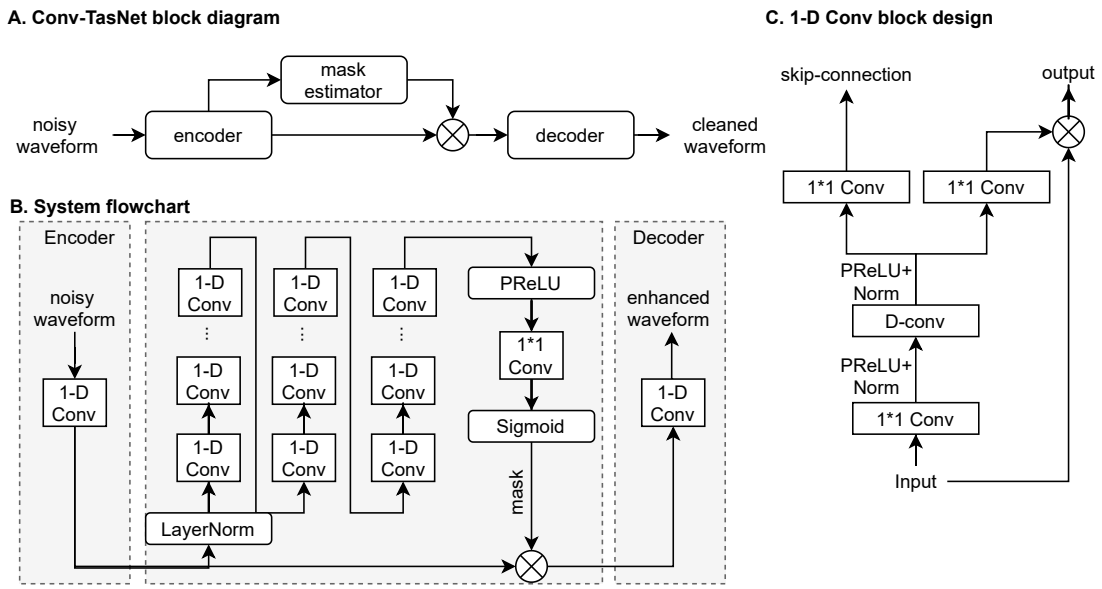


FIGURE 2.8: The block diagrams of (A) the workflow of the Conv-TasNet, (B) the structure of the Conv-TasNet, and (C) the structure of 1-D convolutional block. The encoder extracts a high-dimensional acoustic representation from noisy waveforms and a separation module predicts a mask for noisy acoustic representation. The decoder reconstructs cleaned waveforms from the representations after the mask. Different dilation factors are presented by different colors in the Conv-TasNet. This diagram is re-drawn from the prior work [1].

(with 512 filters and 1×1 kernel size) determines the input channels and the second 1×1 CNN (with 128 filters and 1×1 kernel size) adjusts the output channels from the dilated convolutional layer (with 512 filters and 1*3 kernel size).

Eight TCNs are formed as a group and each group is repeated for 3 times in the mask estimation module. In each group, we increase the dilation factors of the depth-wise convolutions in the 8 TCNs as $[2^0, \ldots, 2^{b-1}]$. To keep the dimension of the estimated mask M in consistent with the encoder representations, one 1×1 CNN (with 512 filters and 1×1 kernel size) is applied with a sigmoid activation function to ensure that the estimated mask M ranges within [0, 1].

In the training stage, the initial learning rate is set to 1e-3 and ADAM is utilized as the optimizer. The objective function is the scale invariant source-to-noise ratio (SI-SNR), which can acquire better speech quality in this framework. The SI-SNR is defined as:

$$\rho(\hat{s}, s) = 10 \log_{10}\left(\frac{\left\|\frac{\langle \hat{s}, s\rangle}{\langle s, s\rangle} s\right\|^2}{\left\|\frac{\langle \hat{s}, s\rangle}{\langle s, s\rangle} s - \hat{s}\right\|^2}\right) \tag{2.21}$$

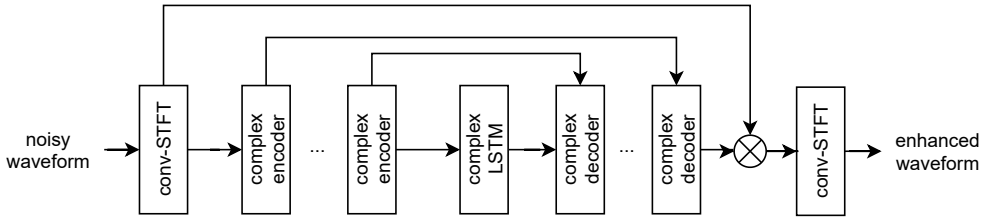


FIGURE 2.9: The block diagrams of DCCRN network. This diagram is re-drawn from the prior work [7].

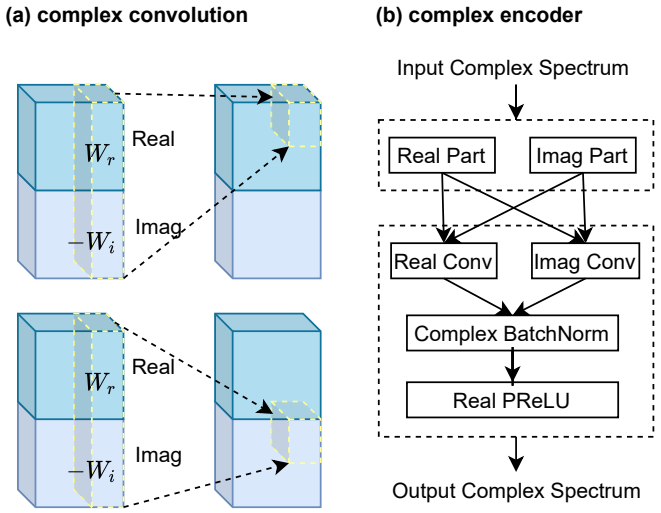


FIGURE 2.10: The block diagrams of a) complex convolution and b) complex encoder [7].

where $\hat{s}$ is the enhanced signals and s is clean labels of noisy inputs. $\langle , \rangle$ conducts the inner product. The signals $\hat{s}$ and s are conducted zero-normalization to ensure scale invariance.

2.3.3 Complex-domain approach

In this section, we will introduce a deep complex convolution recurrent network as the representative complex-domain approach [7].

Prior work [78] recently proposed one encoder with two decoders structure to learn the real and imaginary parts of complex STFT spectrograms from the noisy input to clean labels. Different from the traditional magnitude-only target, modeling magnitude and phase together improves the speech performances significantly. However, this work treats real and imaginary parts as two separate inputs. The convolution operation is real-valued, which does not fully utilize the information

of complex spectrum. In this way, the knowledge between the real and imaginary parts are not learned by the network.

To alleviate such problem, the complex convolutional layer and the complex batch normalization layer are introduced in the encoder/decoder of the DCCRN method. Likewise, complex LSTM replaces the real-valued LSTM. Therefore, the complex network could learn the correlated knowledge between magnitude and phase. Specifically, the encoder extracts high-level acoustic features from the input and the decoder reconstructs the enhanced features back to the original input. The Conv2d block of encoder/decoder consists of a convolution/deconvolution layer with a batch normalization operation and ReLU active function. The complex batch normalization and PReLU follow the implementation of prior work. Skip-connection concentrates the encoder and decoder to speed up the training process. The complex-valued convolutional filter W is defined as,

$$W = W_r + jW_i \tag{2.22}$$

where W_r and W_i denotes the real-valued matrices of a complex convolution kernel, respectively. Therefore, the input complex matrix is defined as,

$$X = X_r + jX_i \tag{2.23}$$

Thus, the complex output Y produced by the complex convolution operation $X \bigotimes W$ is defined as:

$$F_{out} = (X_r \times W_r - X_i \times W_i) + j(X_r \times W_i + X_i \times W_r) \tag{2.24}$$

where F_{out} is the output feature. Likewise, complex LSTM output F_{out} can be defined with the real and imaginary parts of the complex input X_r and X_i, :

$$\begin{aligned} F_{rr} &= LSTM_r(X_r); \\ F_{ir} &= LSTM_r(X_i); \\ F_{ri} &= LSTM_i(X_r); \\ F_{ii} &= LSTM_i(X_i); \\ F_{out} &= (F_{rr} - F_{ii}) + j(F_{ri} + F_{ir}) \end{aligned} \tag{2.25}$$

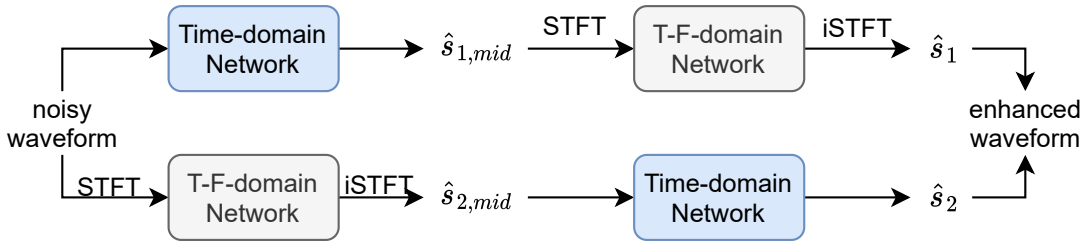


FIGURE 2.11: The block diagram of the multi-domain processing via hybrid denoising (MDPHD) network. The networks in same color share the parameters with each other. The frequency-domain network extracts acoustic features directly from waveforms via the short time Fourier transform (STFT). The enhanced speech is reconstructed to time-domain signals via the inverse short time Fourier transform (iSTFT). This diagram is re-drawn from the prior work [3].

where $LSTM_r$ and $LSTM_i$ are two traditional LSTMs of real part and imaginary part.

2.3.4 Multi-domain approach

In this section, we will introduce a multi-domain processing via hybrid denoising network (MDPHD) as the representative multi-domain approach [3].

Specifically, the MDPHD method consists of two branching workflow: the time-domain branch and the frequency-domain branch, as shown in Figure 2.11. The time-domain branch follows the implementation of Conv-TasNet [1] structure, which proposes 1-D dilated convolutional layer to learn long-range knowledge of inputs. For the frequency-domain network, a U-Net structure [77] based on 2-D convolutional layers is employed [79, 80] to learn an ideal ratio mask (IRM) for noisy inputs. The noises can be removed by multiplying the estimated mask to the noisy spectrogram in frequency space.

At the training stage, the energy-conserving loss function for each of the two branches is utilized, which take consideration of speech and noise signals together. Suppose that the noisy input y is mixed by clean speech s and noise d, the enhanced speech is denoted by $\hat{s}$. Thus, the loss function is formulated as:

$$L(y, s, d, \hat{s}) = \|s - \hat{s}\|_1 + \|d - \hat{d}_1\| \tag{2.26}$$

where $\hat{d} = y - \hat{s}$ is the predicted noise signal. $\|\cdot\|_1$ is l_1 norm.

To balance the contribution of two branches, the loss $L(x, s, n, \hat{s}_{i,mid})$ is introduced at the intermediate conjunction. $\hat{s}_{i,mid}$ is the output of the former network. In addition, the entire model switches the sequential order of each component for fully training. As the time and frequency domain networks are hybridized in a cascaded way, the final objective of the hybrid model is defined as,

$$min_\theta \sum_{i=1,2} L(y, s, d, \hat{s}_{i,mid}) + \sum_{i=1,2} L(y, s, d, \hat{s}_i) \tag{2.27}$$

where θ are parameters of the whole network.

At the testing stage, either a single path result or the averaged results can be used. In this work, the averaged output shows the best performance.

2.4 Limitations of current speech enhancement technologies

In this section, we will summarize the limitations of the current deep-learning-based speech enhancement technologies including frequency-domain approaches, time-domain approaches, complex-domain approaches and multi-domain approaches.

Limitations of frequency-domain approaches. Frequency-domain approaches usually transform time-domain waveforms to frequency-domain features via the STFT, which can produce stable acoustic features for subsequent masking predicting. However, some studies [1, 28] find that frequency-domain methods have several unavoidable limitations. First, STFT is a generic signal transformation for all speech tasks, which might not be the most suitable feature transformation approach for speech enhancement. Second, frequency-domain approaches usually ignore phase information when extracting the frequency-domain spectrum. Although methods for phase reconstruction can be applied to alleviate this issue [31–35], the erroneous estimation of the phase introduces an upper bound on the accuracy of the reconstructed audio [29, 30]. This issue is evident by imperfect reconstruction even when the ideal clean magnitude spectrograms are utilized.

Limitations of time-domain approaches. Due to the unavoidable limitations of frequency-domain techniques, time-domain approaches are proposed to avoid

the phase manipulation issue by operating directly on the noisy waveform and eliminating an explicit STFT, so that the whole encoder-decoder-like structure can be optimized while omit the phase estimation. However, as the time-domain encoder and decoder are fully learned from the training data, they are sensitive to varying testing environments [16, 48, 49]. Recent studies also [50–52] observed that time-domain speech separation/enhancement frameworks might not always yield good performances facing varying testing scenarios. The reason for this problem is in two-folds. First, the convolutional filters in the time-domain speech encoder fully rely on training data. They cannot always yield a good decomposition of the input speech compared with the fixed STFT algorithm when faced with various test data. Second, the end-to-end training strategy cannot guarantee that the speech encoder and decoder are well trained through this process, which might also cause the sensitivity of time-domain speech separation/enhancement frameworks.

Limitations of complex-domain approaches. Complex-domain technologies process the real and imaginary values of the spectrum features via deep learning networks, therefore, phase information can also be predicted accurately. However, as real and imaginary are processed separately, the models usually require more memory during processing than the single network.

Limitations of multi-domain approaches. As time-domain techniques are sensitive to varying testing environments, multi-domain approaches are proposed to combine frequency-domain approaches and time-domain approaches to improve the robustness of the models. Current studies take a combination of two ways: cascade combination [3] and parallel combination [4]. However, as two domain network are combined, the resulting multi-domain models are usually huge, leading to increases in training time. Current performance of such methods remains sub-optimal. Smarter combination of frequency-domain approaches and time-domain approaches are still being explored.

In addition to the above inherent limitations of frequency-/time-/multi-domain frameworks, the mismatch problem is the common issue that speech enhancement techniques are facing as learning-based speech enhancement approaches usually assume that the training and testing data have the same probability distribution. However, practical scenarios often fail to meet this assumption. Therefore, speech enhancement performance may degrade significantly in face of mismatched scenarios at run-time. Many factors cause the mismatch problem in speech enhancement,

such as unseen speakers, unseen accents, *unseen noises in the testset, the channel effect and the sensitive time-domain encoder/decoder*. This research focuses on solving the problem of mismatch caused by the last three factors in speech enhancement models.

2.5 Summary

This chapter first introduces the basic information of speech enhancement, including the definition, the commonly-used corpus and the evaluation metrics. We then briefly describe the two representative approaches for speech enhancement models used prior to the deep learning era (i.e., spectral subtraction approach and minimum mean square error based approach) and the four deep-learning-based models (i.e., frequency-domain approach, time-domain approach, complex-domain and multi-domain approach). Finally, we analyse the limitations of the current deep-learning-based approaches and point out which problem this thesis will focus on.

Chapter 3

Speech Enhancement with Adversarial Training

This chapter[1] focuses on alleviating the mismatch problem caused by unseen noises in the testset. Our solutions consider two practical scenarios: with the target-domain data and without the target domain data. One proposed approach utilizes the adversarial training strategy for speech enhancement to learn the domain-agnostic features under the condition with the target-domain data. The other proposed approach attempts to learn the noise-agnostic features with the adversarial training under the condition without the target-domain data.

Section 3.1 introduces the motivation of two proposed frameworks and related studies. Section 3.2 presents the proposed domain adversarial training approach for speech enhancement (SE-DAT) under the condition that only the noisy target-domain data is available. Section 3.3 describes the proposed disentangled feature learning approach with the adversarial training strategy (NAT-SE) under the condition that no target-domain data is available. Section 3.4 concludes this chapter.

3.1 Motivation

Speech enhancement techniques aim to reduce abundant background noise to speech signals for better speech intelligibility and quality. As discussed in Chapter 1.1, for

[1]The works in this chapter have been published in [81, 82]

such machine learning tasks, the training and testing data are usually assumed to have the same probability distribution. However, practical scenarios often fail to meet this assumption, which is commonly caused by the unseen data in the testset. As a result, speech enhancement performance may degrade significantly in face of such unseen noises at run-time.

To address the mismatch problem caused by unseen noises in the testset, some techniques have been studied to adapt or transfer the speech enhancement model to unseen conditions, such as domain adaptation [81, 83, 84] and teacher-student learning [85–87].

Domain adaptation techniques aims to adapt a model trained under one training condition (i.e., the source domain) towards another testing condition (i.e., the target domain). For example, the study in [46] suggests adapting the last layers of pre-trained speech enhancement generative adversarial network (SEGAN) with the dataset of new language and noise to reduce the mismatch between different languages and noise, but this technique asks for clean-noisy parallel speech data in target domain that are not always available in practice.

Teacher-student learning techniques [85–87] transfer the invariant-knowledge from source domain to target domain, or teacher to student. The student model, therefore, is taught to keep the invariant-knowledge from the teacher model and adapt to the variations in the target domain. However, such teacher-student learning based techniques also require the auxiliary information like the transcripts, which might be not always available in the speech enhancement corpus.

In this chapter, we aim to alleviate the mismatch problem caused by unseen noises under two practical scenarios: with the noisy target-domain data and without any target-domain data.

- If only noisy target-domain data is available, we propose a domain transfer approach by adapting the enhancement net without the need of clean-noisy parallel speech data in the target domain. In our scenarios, the training data in source domain consist of clean-noisy speech pairs, but those in the target domain only consist of noisy speech. We propose a complete pipeline to overcome the mismatch across domain, that we call domain adversarial training approach to speech enhancement (SE-DAT).

- If no target-domain data is available, we propose a network architecture with noise adversarial training (NAT-SE) to derive noise-agnostic feature representations. The network seeks to address the unseen noise problem in the target domain without the need of any target domain data.

The two proposed architectures will be introduced in the following Section 3.2 and Section 3.3, respectively.

3.2 SE-DAT with target-domain data

Recently in image processing, the domain adaptation technique [88] is utilized to adapt features with a domain discriminator structure via domain adversarial training (DAT) in face of test data in the new domain. In speech and speaker recognition, it was used to adapt acoustic models or produce speaker-invariant features to overcome the mismatch between training and testing [84, 89, 90].

Inspired by prior study in DAT, we propose a domain adversarial training approach for speech enhancement. It has the following main advantages:

- SE-DAT can be adapted to target domain with only noisy speech data, without the need of clean-noisy speech pairs.
- SE-DAT architecture is concise and requires no deep structure like feature extractor in the work [83] to learn adapted feature representation.
- We also introduce the dynamic features [91, 92] into SE-DAT, which takes the temporal context of features into consideration to ensure the continuity of the enhanced speech [93, 94].

3.2.1 The proposed architecture

Specifically, we assume the model learns the mapping between noisy speech sample $x \in X$ and its corresponding clean sample $y \in Y$. y and x form a clean-noisy speech pair. We also assume that the noisy sample x and its clean sample y belong to a distribution $\mathcal{S}(x, y)$, also called source domain. Suppose that we now have

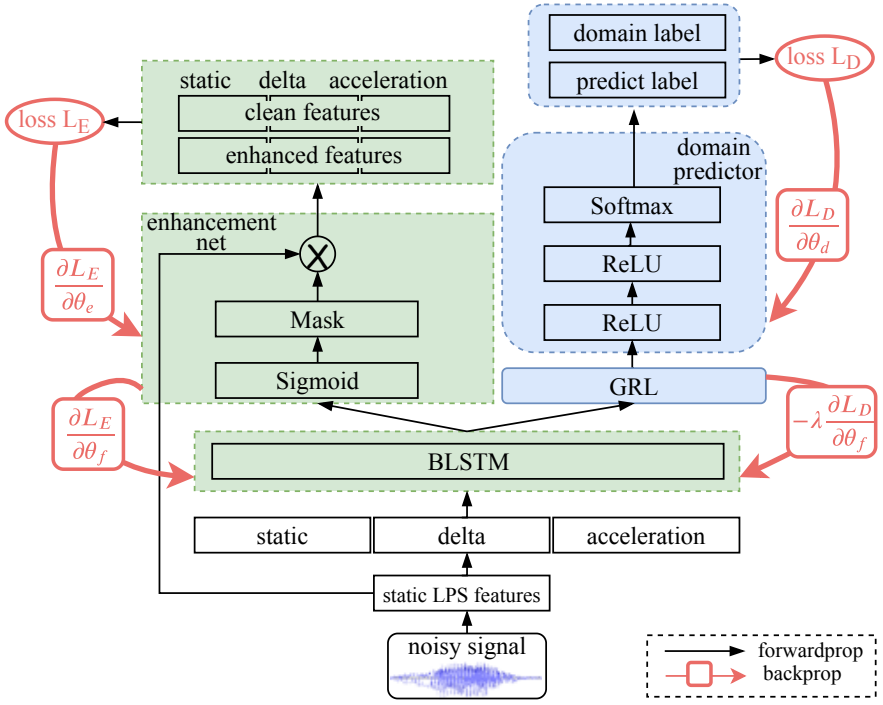


FIGURE 3.1: SE-DAT includes two parts, an enhancement net E (green) that generates the enhanced speech and a domain predictor D (blue) that distinguishes between domains the input comes from. The two parts are jointly trained to minimize the loss of the enhancement net L_E and to maximize the loss of the domain predictor L_D at the same time through a GRL.

some noisy speech samples in the new domain without the corresponding clean speech samples, we hope to adapt the model so that it works both in the source domain and the new domain. The unpaired dataset is assumed to belong to the other distribution $\mathcal{T}(x, y)$, also called target domain. Finally, we assign the binary domain label, $d \in [0, 1]$, to each noisy sample at the training stage to indicate which domain the noisy samples come from. We illustrate the proposed technique in Figure 3.1.

3.2.1.1 Dynamic features

In the training process, the training speech data from both the source domain and the target domain are extracted with a shifting window into static log-power-spectrum (LPS) features, that is also called the static feature. LPS is selected here as it can convert a linear power spectrum into a logarithmic space. A frame of speech is represented by a vector of static features. A limitation of using only LPS

features is that each frame is represented independently and we cannot guarantee that the produced frame sequence is smooth and sounds natural. Hence, we introduce the dynamic features [91, 92] that take the temporal context of features into consideration. In this work, the dynamic features are the derivatives of the LPS features, including delta features (first-order time derivatives) and acceleration features (second-order time derivatives). We can approximate the delta features and acceleration features as follows:

$$f_D(t) = \frac{\sum_{l=1}^{L} l * (f_S(t+l) - f_S(t-l))}{\sum_{l=1}^{L} 2l^2} \tag{3.1}$$

where $f_S(t)$ and $f_D(t)$ are the static feature and the delta feature respectively at frame t. L is the order of computing the derivatives and is set to 2 in this study. The acceleration features denoted as f_A are obtained by applying equation (3.1) on the delta features f_D. The original LPS feature, delta feature, and acceleration feature jointly form a new feature $F = [f_S, f_D, f_A]$ for a speech frame.

3.2.1.2 The enhancement net

The enhancement net E aims to map input noisy speech to clean speech by estimating a mask, where one bidirectional long short-term memory (BLSTM) layer produces the adapted representations v for input feature frame $F = [f_S, f_D, f_A]$. Such representations v is used by two nets: the enhancement net E and the domain predictor D. In the enhancement net E, suppose that the representations v_i of input sample x_i arrives from the source domain, we take the dot product between the static LPS feature f_S and its estimated mask. We then take the enhanced static feature $\hat{y}_S$ to obtain its dynamic features $\hat{y}_D$ and $\hat{y}_A$ according to equation (3.1). Finally, we compute the spectrum approximation loss between the enhanced feature frame $\hat{y} = [\hat{y}_S, \hat{y}_D, \hat{y}_A]$ and the corresponding clean feature frame.

The spectrum approximation loss for enhancement net E [92] is given as follows:

$$\begin{aligned} L_E(\theta_f, \theta_e) = \|\hat{y}_S - y_S\|_F^2 + w_D\|\hat{y}_D - y_D\|_F^2 \\ + w_A\|\hat{y}_A - y_A\|_F^2 \end{aligned} \tag{3.2}$$

where θ_f and θ_e are parameters of the BLSTM layer and the rest enhancement net respectively. $\|\cdot\|_F$ is the Frobenius norm. w_D and w_A are the weights of cost

contributed by the delta and acceleration features. Besides, if the representations v_i for input sample x_i arrives from the target domain (without the paired clean-noisy speech), we don't calculate the loss L_E for this input noisy sample due to no clean reference.

3.2.1.3 The domain predictor

SE-DAT aims to overcomes the mismatch between the source and target domain without the need of clean-noisy parallel data, which is achieved by the domain predictor. In domain predictor D, we set the i-th domain label as d_i for the representation v_i to indicate where v_i comes from. If v_i comes from the source domain, d_i is set to 0 (if $v_i \sim \mathcal{S}(v)$, set $d_i = 0$), otherwise d_i is set to 1 (if $v_i \sim \mathcal{T}(v)$, set $d_i = 1$). The cross-entropy loss for domain predictor D is defined as:

$$\begin{aligned}\mathcal{L}_D(\theta_f, \theta_d) = -\frac{1}{N}\sum_{i=1}^{N}[d_i \log P(v_i \in \mathcal{S}(v)) \\ + (1 - d_i) \log P(v_i \in \mathcal{T}(v))]\end{aligned} \tag{3.3}$$

where θ_f and θ_d are parameters of the BLSTM layer and the domain predictor D respectively. N is the number of input training samples.

We now jointly train the two parts: the enhancement net E and the domain predictor D for 1) seeking the parameters θ_f to maximize the loss of the domain predictor D, 2) simultaneously seeking the parameters θ_d to minimize the loss of domain predictor D, and 3) seeking θ_e to minimize the loss of the enhancement net E. Such optimization can be achieved by the gradient reversal layer (GRL). The role of GRL is an identity transform during the forward propagation. During the backpropagation, the GRL multiplies the gradient from the domain predictor D by $-\lambda$ and then passes it to the BLSTM layer. The whole cost function of the SE-DAT is formulated below:

$$\mathcal{L}(\theta_f, \theta_e, \theta_d) = \mathcal{L}_E(\theta_f, \theta_e) - \lambda \mathcal{L}_D(\theta_f, \theta_d) \tag{3.4}$$

where λ is the gradient reversal coefficient that controls the trade-off between two objectives during training. λ is defined as:

$$\lambda = \frac{2}{1 + \exp\left(-10 * \frac{j+k*J}{K*J}\right)} - 1 \tag{3.5}$$

where j denotes the index of current batch and J is the total number of batches. k presents the index of current epoch and K is the total number of epochs. In this way, standard stochastic gradient solvers (SGD) can be applied for the search of the best parameters $(\theta_f, \theta_e, \theta_d)$ as follows:

$$\begin{aligned}
\theta_f &\leftarrow \theta_f - \mu\left(\frac{\partial L_E}{\partial \theta_f} - \lambda \frac{\partial L_D}{\partial \theta_f}\right) \\
\theta_e &\leftarrow \theta_e - \mu \frac{\partial L_E}{\partial \theta_e} \\
\theta_d &\leftarrow \theta_d - \mu \frac{\partial L_D}{\partial \theta_d}
\end{aligned} \tag{3.6}$$

where μ is the learning rate.

At the testing stage, only the noisy speech is enhanced by the enhancement net E, while the domain predictor D is discarded.

3.2.2 Experiments and results

We would like to validate the proposed SE-DAT by adapting the enhancement net from source domain to target domain.

3.2.2.1 Database

We conduct experiments of SE-DAT on two corpora: one is CHiME-4 dataset [95]; the other is the dataset released by Cassia Valentini-Botinhao [96], which is same in SEGAN [42] and Wave-U-Net [39], referred to as VCTK dataset hereafter. We use CHiME-4 dataset as source domain data and VCTK dataset as target domain data in order to validate SE-DAT in reducing the mismatch across domains.

Source domain: CHiME-4 dataset

In the CHiME-4 dataset, the simulated data are generated by artificially mixing clean speech data with noisy backgrounds of four types, i.e. cafe, bus, street, and

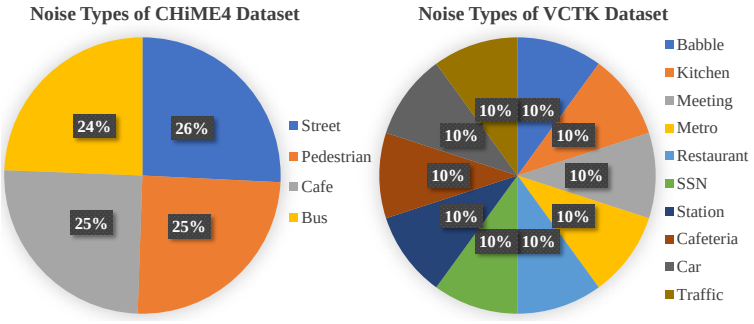


FIGURE 3.2: The statistics of noise types of CHiME4 dataset (left) and VCTK dataset (right).

pedestrian area. We use the simulated training set (7,128 utterances) and simulated development set (1,600 utterances) as source domain data.

Target domain: VCTK dataset

The VCTK dataset considers a total of 40 different conditions [96]. Specifically, it has 10 types of noise, including 2 artificial noises and 8 from the Demand database [54]. These noises are mixed with clean data by 4 signal-to-noise ratio: 15dB, 10dB, 5dB, and 0dB. There are 14 male and 14 female training speakers. We use the VCTK dataset at a ratio of 9:1 as the training set (10,415 utterances) and development set (1,157 utterances). Likewise, the test set considers a total of 20 different conditions [96]. It includes 5 types of noise mixed by 4 SNR: 17.5dB, 12.5dB, 7.5dB, and 2.5dB, with 1 male and 1 female test speakers.

As shown in Figure 3.2, the conditions of the CHiME-4 dataset and the VCTK dataset are different in noise types. Besides, the training speakers and SNR are totally different, which fits our purpose: evaluating the effectiveness of SE-DAT in reducing the mismatch between two different domains. To show the effectiveness of DAT, we use VCTK dataset in the noisy target domain without the need of its corresponding clean speech.

3.2.2.2 Experimental setup

The two datasets are sampled at 16 kHz sampling rate and 16 bits/sample. We applied 512-point STFT to extract LPS, the delta features and acceleration features. One BLSTM layer is used with 512 units, which is followed by one feed-forward layer of 257 logistic units with sigmoid activation in enhancement net E. The

domain predictor D consists of three feed-forward layers with two ReLU activations and one softmax activation. The learning rate μ equals 0.001, and the batch size is 32. The weights for delta features w_D and for acceleration features w_A are empirically set to 4.5 and 10.0 respectively [93, 97]. Early stop and learning rate adjustment strategy are also adopted in the experiments. The start halving improvement, halving factor and the end halving improvement are 0.003, 0.5 and 0.001 respectively.

To evaluate SE-DAT, two models were trained using aforementioned datasets:

- SE-DAT-0: The model, with λ set to 0, is trained only on source domain CHiME-4 data (with clean-noisy parallel data) and tested on target domain VCTK test set. This model serves as the reference baseline for testing, where model adaptation is no attempted.
- SE-DAT: SE-DAT is trained by both source domain CHiME-4 data (with clean-noisy parallel data) and target domain VCTK data (noisy speech without clean speech counterpart) to verify the effectiveness, which attempts to use the noisy target domain data to overcome the mismatch across domains.

3.2.2.3 Results

In this work, we compute the following objective measures. All metrics compare the enhanced signal with the clean reference on the VCTK test set (824 utterances), using the tookit in [73].

- PESQ: Perceptual evaluation of speech quality, ranging from -0.5 to 4.5, which is calculated by the wide-band version recommended in ITU-T P.862.2 [98].
- CSIG: Mean opinion score (MOS) prediction of the signal distortion compared with the speech signal [99], ranging from from 1 to 5.
- CBAK: MOS prediction of background noise compared with the speech signal [99], ranging from from 1 to 5.
- COVL: MOS prediction of the overall effect [99], ranging from from 1 to 5.

TABLE 3.1: Comparisons with SE-DAT-0 and SE-DAT in terms of the PESQ, CSIG, CBAK, COVL and SSNR scores on VCTK test set. "Zero-effort" means that we use the untreated noisy speech of VCTK test set. Higher scores are better for all metrics.

Method	PESQ	CSIG	CBAK	COVL	SSNR
Zero-effort	1.97	3.35	2.44	2.63	1.68
SE-DAT-0	2.12	3.38	2.46	2.66	1.76
SE-DAT	**2.26**	**3.72**	**2.77**	**2.98**	**4.11**

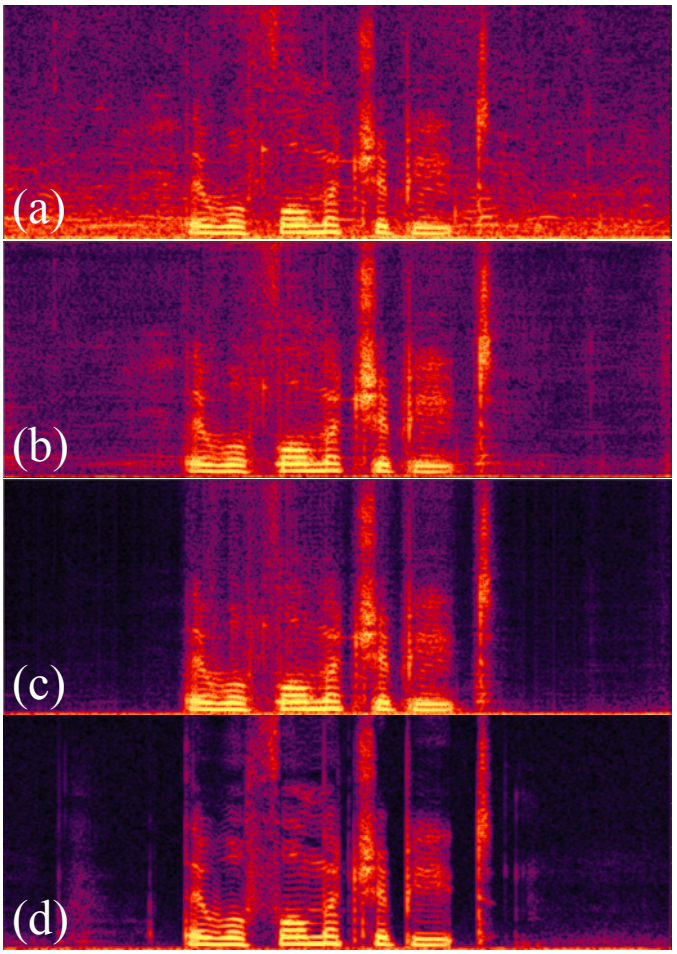


FIGURE 3.3: Comparisons of spectra. (a) denotes the spectrum of the noisy speech and (b) is the spectrum of the corresponding enhanced speech by SE-DAT-0. (c) represents the spectrum of the corresponding enhanced speech by SE-DAT and (d) is the spectrum of the corresponding clean speech.

- SSNR: Segmental SNR [100] (from 0 to ∞).

Effect of the SE-DAT:

As shown in Table 3.1, we note that SE-DAT-0 trained on CHiME4 simulated training set alone does not perform well on the VCTK test set in the new domain.

TABLE 3.2: Training details of different methods on VCTK dataset.

Method	Training set	Clean for supervision	Feature domain	Test set
SEGAN [46]	VCTK set	Yes	time domain	VCTK test set
CNN-GAN [75]	VCTK set	Yes	frequency domain	VCTK test set
Wave-U-NET [39]	VCTK set	Yes	time domain	VCTK test set
SE-DAT	CHiME4 simu set for source domain VCTK set for target domain	Yes for source domain No for target domain	frequency domain	VCTK test set

TABLE 3.3: Comparisons with different methods in terms of the PESQ, CSIG, CBAK, COVL and SSNR scores on VCTK test set. "Zero-effort" means that we use the untreated noisy speech of VCTK test set.

Method	Training	PESQ	CSIG	CBAK	COVL	SSNR
Zero-effort	–	1.97	3.35	2.44	2.63	1.68
Wiener [101]	–	2.22	3.23	2.68	2.67	5.07
SEGAN [46]	supervised	2.16	3.48	2.94	2.80	7.73
CNN-SEGAN [75]	supervised	2.34	3.55	2.95	2.92	–
Wave-U-Net [39]	supervised	**2.40**	3.52	**3.24**	2.96	**9.97**
SE-DAT	unsupervised	2.26	**3.72**	2.77	**2.98**	4.11

With the domain mismatch, we observe that the performance of SE-DAT-0 is almost same as the noisy speech without enhancement. By applying DAT, we observe that the proposed SE-DAT approach drastically improves all the performance when we train only on noisy target domain data.

To further showcase the ability of SE-DAT, a speech utterance (spectrum) from VCTK test set is shown in Figure 3.3. The original noisy speech is shown in (a) and the corresponding clean speech is in (d). We observe in (b) that SE-DAT-0 cannot reduce the noise effectively in unseen speech of the new domain and most of the noise components still remain. By contrast, despite training without the clean speech utterances in the new domain, the proposed SE-DAT still can significantly remove the noise components as shown in (c).

Benchmark against the baselines:

We further compare the proposed SE-DAT with some recent methods conducted on VCTK dataset although this comparison is not fair. As shown in Table 3.2, the methods like SEGAN, CNN-GAN, and Wave-U-Net are all trained using clean-noisy paired VCTK speech to supervise the learning of the network without mismatch problem. We are glad to see that the proposed SE-DAT transferred the knowledge from the source domain to the target domain without supervision information from clean speech in the target domain as reference during training. In addition, SEGAN and CNN-GAN directly extract the features from time domain, which means there is no phase problem. CNN-GAN and the proposed SE-DAT

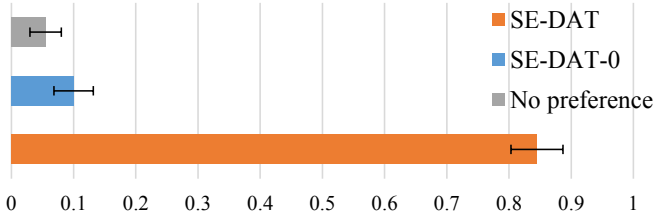


FIGURE 3.4: Results of the quality preference test with 95% confidence intervals for different methods.

use the spectrum features through STFT and re-use the phase of noisy speech. As shown in Table 3.3, despite the unfair conditions, the proposed SE-DAT still performs better in CSIG and COVL, which means it produces less speech distortion and achieves a better overall quality.

Subjective evaluation:
The AB preference test was conducted to assess the subjective perceptual quality of the enhanced speech. In the AB preference test, each paired samples A and B were randomly selected from the proposed SE-DAT model and the SE-DAT-0 model. 10 subjects participated in the preference test. Each listener was asked to choose the sample with better quality from each pair. We encourage subjects to wear headphones. The samples are randomly presented without framing sounds. Repetition are allowed and forced choice is applied. The subjective results of quality preference test are presented in Figure 3.4. The results suggest that the speech quality of SE-DAT significantly outperforms that of SE-DAT-0.

3.3 NAT-SE without target-domain data

To alleviate the mismatch problem caused by unseen noises without target-domain data, we propose a disentangled feature learning framework for speech enhancement (NAT-SE) with adversarial training under the condition without target-domain data. The NAT speech enhancement, or NAT-SE in short, network adopts an encoder-masking-decoder architecture as the generator and a noise type classifier with a gradient reversal layer (GRL) as the disentangler. The disentangler removes the unspecified noise factors from the feature representation, and the decoder ensures that the feature representation maintains the clean speech content. Specifically, the noise factor is encoded via the classification task to distinguish

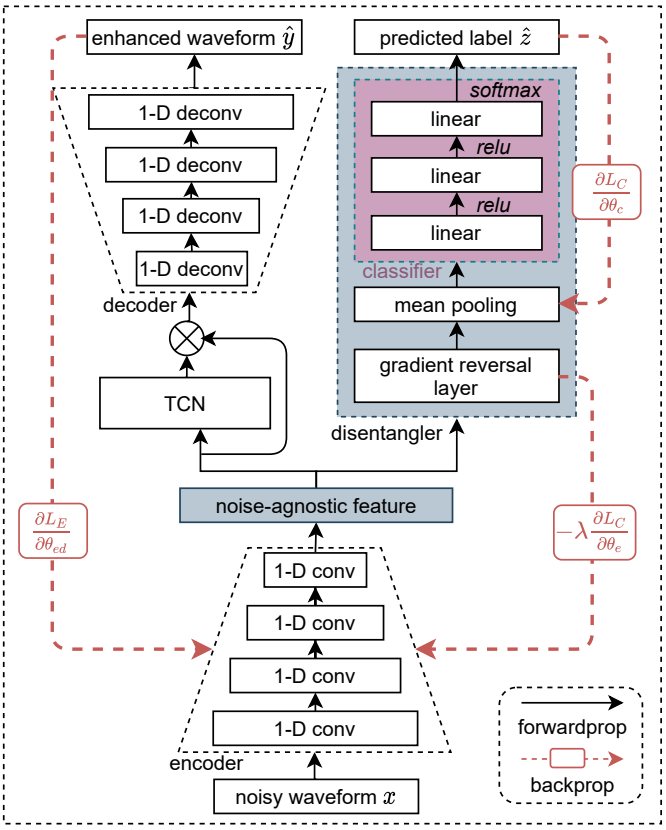


FIGURE 3.5: Block diagram of the proposed NAT-SE. $\otimes$ is the element-wise multiplication. L_E denotes the enhancement loss of SI-SDR and L_C is the cross-entropy loss of the noise classifier. λ is the positive gradient reversal coefficient.

various type of noises. Such noise factor is then removed from the encoded feature representations with the gradient reversal layer.

3.3.1 The proposed architecture

The proposed noise adversarial training speech enhancement (NAT-SE) framework consists of four modules: encoder, temporal convolutional network (TCN) based mask estimator, disentangler and decoder, as illustrated in Figure 3.5.

The convolutional encoder, consisting of several 1-D convolutional layers, extracts acoustic features from the noisy waveforms x, which is widely used in enhancement and separation tasks [1, 102]. We firstly employ filter length (2 *samples*) with a stride same as the filter length in each layer of the encoder. The number of filters in the last encoder layer is set to 512 and is halved successively in early layers. As the stride is set as 2 in the 1-D convolutional layers, the temporal dimension of

features is halved after a 1-D layer upwards in the multi-layer encoder, and doubled after a 1-D layer upwards in the multi-layer decoder as in Figure 3.5.

Then, the disentangler module attempts to predict the type of noises in encoded representations. With the GRL, the encoder, therefore, is updated in opposition of classification task and learn the noise-agnostic features. Meanwhile, the TCN is utilized to predict the mask for the noise-agnostic features to filter out the residual noise. Finally, a transposed-convolutional decoder reconstructs the speech waveform from the enhanced features. The details of the disentangler and the TCN-based mask estimator are described as follows.

3.3.1.1 Disentangler

The disentangler module (DM) is designed to predict the types of noises in the encoded representations via the noise classifier in the forward propagation and force the encoder to generate the noise-agnostic features via the GRL in the back-propagation. In the training process, no target-domain data is required. Specifically, the disentangler module consists of a GRL, a mean pooling layer and a classifier. The role of GRL is an identity transform during the forward propagation. During the back-propagation, the GRL multiplies the gradient from the noise classifier C by -λ and then passes it to the encoder, where λ is the positive gradient reversal coefficient. The classifier consists of three linear layers and the third layer uses the softmax activation function to classify the type of noises. We adopt the cross-entropy loss $L_C(\hat{z}, z)$ for the noise classifier, where $\hat{z}$ and z are the predicted noise types and the real noise labels.

3.3.1.2 TCN-based mask estimator

The TCN-based mask is designed to suppress the additive noise in encoded representations. Similar to Conv-TasNet [1], the mask estimation module consists of a temporal convolutional network (TCN). As shown in Figure 4.3, the encoded representation firstly performs mean and variance normalization on channel dimension, which is scaled by the trainable bias and gain parameters. Then, a 1×1 CNN with $N = 128$ filters adjusts the numbers of input channels. To learn the long-range temporal information of the speech with few parameters, we stack dilated

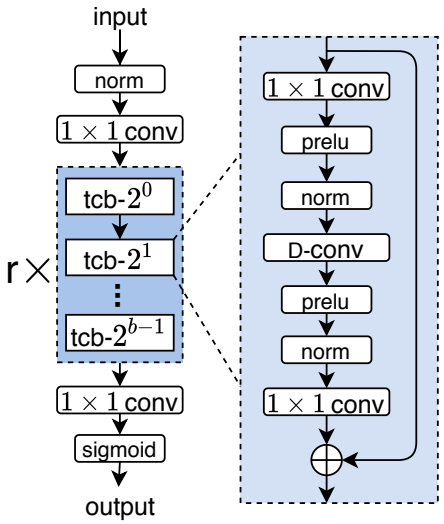


FIGURE 3.6: Block diagram of the temporal convolutional network (TCN). "*tcb*-2^{b-1}" denotes a temporal convolutional block (TCB) with the dilation of 2^{b-1}, where b is the total number of the TCB. "D-conv" is the dilated depthwise convolutional layers stacked in several TCBs to exponentially increase the dilation factors. $\bigoplus$ is the residual connection.

depthwise convolutional layers "D-conv" in several temporal convolutional blocks (TCB) with exponentially increased dilation factor $[2^0, \ldots, 2^{b-1}]$. In this work, $b = 8$ TCBs are formed as a batch and each batch is repeated for $r = 3$ times in the TCN-based mask. To keep the TCN-based mask in a consistent dimension with the input features, one 1×1 CNN (with 512 filters and 1×1 kernel size) is applied with a sigmoid activation function for ensuring that the estimated mask ranges within $[0, 1]$.

3.3.1.3 Adversarial training strategy

NAT-SE adopts the following optimization strategies: 1) seeking θ_{ed}, the parameters of the basic encoder-masking-decoder structure, to minimize the enhancement loss L_E via scale-invariant signal-to-distortion ratio (SI-SDR), 2) simultaneously seeking θ_c, the parameters of the classifier, to minimize the cross-entropy loss of the classification task L_C, 3) seeking θ_e, the parameters of the encoder, to maximize the cross-entropy loss of the classification task L_C. Such optimization is achieved by the adversarial training with the GRL. The whole cost function L is formulated below:

$$L(\theta_{ed}, \theta_c, \theta_e) = L_E(\theta_{ed}) - \lambda L_C(\theta_c, \theta_e) \tag{3.7}$$

where λ is the positive gradient reversal coefficient that controls the trade-off between two objectives during training. The disentangler module is not involved during inference.

3.3.2 Experiments and results

3.3.2.1 Database

We conduct experiments on a publicly available dataset (old version)[2], which consists of 11,572 mono audio samples for training and 824 mono audio samples for testing at sampling rate of 16 kHz [96]. The training dataset consists of 10 noise types, and the test dataset consists of 5 unseen noise types. The unseen noise represents a major source of mismatch between training and test data.

3.3.2.2 Experimental setup

Network configuration:
During the training stage, the noisy waveform is segmented into 1-second frames for batch training. λ is empirically set to 0.5 in the training. The number of filters in the last encoder layer is set to 512 and is halved in early layers. For example, the numbers of filters in the encoder with 4 convolutional layers are set to $64, 128, 256, 512$. The Adam algorithm optimizes the network. The learning rate is set to 0.001, which is halved once the loss increases on the development set for at least 3 epochs. We also apply early stopping scheme as soon as the loss increases on the development set for 20 epochs.

Evaluation metrics:
We report the performances in terms of the following metrics. PESQ [98] stands for perceptual evaluation of the speech quality, ranging from -0.5 to 4.5. Three objective metrics that approximate mean opinion scores (MOSs) [99]: CSIG, CBAK and COVL. They are designed for signal distortion evaluation, noise distortion evaluation, and overall quality evaluation, respectively. Segmental signal-to-noise ratio (SSNR) and signal-to-distortion ratio (SDR) are also conducted for measuring the

[2] `https://datashare.is.ed.ac.uk/handle/10283/1942`

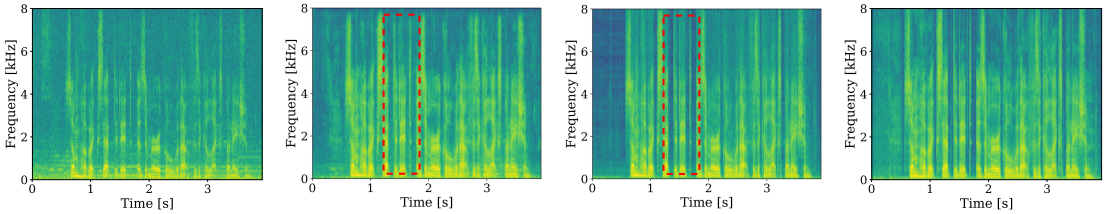


FIGURE 3.7: The spectrograms of a sample (p232_013.wav) in the test set for (a) noisy input, (b) the best baseline Conv-TasNet, (c) enhanced result of NAT-SE and (d) clean signal (ground-truth).

speech quality. Short-time objective intelligibility (STOI) reflects the improvement of speech intelligibility. Higher scores are better for all metrics.

3.3.2.3 Results

Effect of the depth of encoder and decoder:

We first analyse and summarize the performances with different numbers of layers in the encoder and decoder. The disentangler module is not utilized in this experiment. As shown in Table 3.4, the first column denotes the number of layers $(2, 4, 6$ layers) in the encoder and decoder, respectively.

We observe that the performance may not improve as the number of layers increases. We obtain the best performances with 4 layers of encoder and decoder, respectively. The performance with 6 layers drops. This can be explained by the fact that the depth of the encoder is designed increasingly, which causes that the temporal dimension of the encoded representations is reduced to a extreme small scale. We adopt a setting of 4 layers for encoder and decoder hereafter.

TABLE 3.4: PESQ, CSIG, CBAK, COVL, SSNR(dB), SDR(dB) and STOI performances of various depth of encoder and decoder. "#layers" denotes the number of layers for both encoder and decoder. "#Paras" denotes the number of parameters in the model.

#layers	#Paras	PESQ	CSIG	CBAK	COVL	SSNR	SDR	STOI
2	3.92M	2.57	3.83	3.30	3.19	9.22	19.50	0.94
4	**5.10M**	**2.64**	**3.91**	**3.33**	**3.27**	**9.67**	**20.05**	**0.94**
6	5.14M	2.37	3.54	2.26	2.95	7.38	18.45	0.94

TABLE 3.5: PESQ, CSIG, CBAK, COVL, SSNR(dB), SDR(dB) and STOI performances of the disentangler module (DM).

Methods	DM	#Paras	PESQ	CSIG	CBAK	COVL	SSNR	SDR	STOI
WaveUnet [39]	×	10.13M	2.40	3.52	3.24	2.96	9.97	19.82	0.93
	√	10.20M	2.51	3.60	3.32	3.03	10.04	20.21	0.93
NAT-SE	×	5.10M	2.64	3.91	3.33	3.27	9.67	20.05	0.94
	√	5.30M	**2.72**	**3.99**	**3.47**	**3.36**	**10.15**	**20.71**	**0.95**

TABLE 3.6: PESQ, CSIG, CBAK, COVL, SSNR and STOI performances of other competitive methods. Note: the Conv-TasNet [1] utilized the same loss function (SI-SDR) as that in the proposed NAT-SE.

Methods	PESQ	CSIG	CBAK	COVL	SSNR	STOI
Noisy	1.97	3.35	2.44	2.63	1.68	0.91
Wiener [103]	2.22	3.23	2.68	2.67	5.07	–
SEGAN [42]	2.16	3.48	2.94	2.80	7.73	0.93
CNN-GAN [75]	2.34	3.55	2.95	2.92	–	0.93
WaveUnet [39]	2.40	3.52	3.24	2.96	9.97	–
U-Net [104]	2.48	3.65	3.21	3.05	9.34	–
MSE-GAN [105]	2.53	3.80	3.12	3.14	–	0.93
Conv-TasNet [1]	2.57	3.80	3.29	3.18	9.65	–
NAT-SE	**2.72**	**3.99**	**3.47**	**3.36**	**10.15**	**0.95**

Effect of the disentangler module:

We further report the effect of disentangler module (DM) in two network architectures in Table 3.5. We first apply DM to a compact encoder-decoder structure: WaveUnet [39]. We observe that the proposed disentangler module improves the speech quality, such as 4.6% and 2.0% relative improvements in terms of PESQ and SDR. We further compare encoder-masking-decoder structure, that is NAT-SE with and without DM. The proposed NAT-SE achieves 3.0% and 5.0% relative improvements in terms of PESQ and SSNR. We also observe that the parameters of NAT-SE are not increased significantly. Learning the disentangled feature representations has improved the performances in face of unseen noises on test dataset.

NAT-SE vs. other competitive methods:

Table 3.6 summarizes the comparison between the proposed NAT-SE and other competitive techniques in terms of PESQ, CSIG, CBAK, COVL, SSNR and STOI. We observe that the proposed NAT-SE obtained the best performances. Comparing with the Conv-TasNet method, the NAT-SE achieves 5.8% and 5.2% relative improvements in terms of PESQ and SSNR.

To further show the contribution of the NAT-SE approach, we illustrate the magnitude spectrum of an example as shown in Figure 5.4. We can see that the NAT-SE

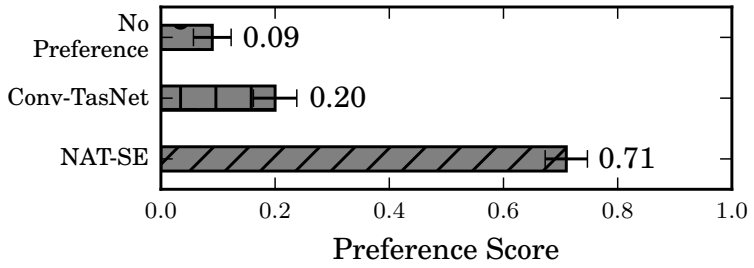


FIGURE 3.8: The result of A/B preference test for the enhanced speech between the proposed NAT-SE and the best baseline Conv-TasNet.

can produce more clear spectrum under mismatched conditions.

Subjective evaluation:

Since the Conv-TasNet presents the best baseline performances in the objective evaluation as shown in Table 3.6, we only conduct an A/B preference test between the Conv-TasNet and the proposed NAT-SE to evaluate the signal quality and intelligibility by subject listening. We randomly selected 20 pairs of listening examples and invited 10 subjects to choose their preference. We show the A/B preference results in Figure 5.8. The results show that 71% listeners preferred the proposed NAT-SE to the best baseline Conv-TasNet, whose preference score is 20%, because there is less mismatch in the proposed NAT-SE.

3.4 Conclusion

In this chapter, we first propose a domain adversarial training technique to speech enhancement (SE-DAT) to overcome the mismatch across domains and provide a solution for speech denoising to the scenario where we don't have clean-noisy parallel data in the new domain. SE-DAT achieves significant improvement on VCTK dataset compared with the model where no effort is made to overcome the mismatch. SE-DAT also delivers voice quality comparable with other supervised learning techniques that require clean-noisy parallel data.

Then, we proposed a noise adversarial training framework for speech enhancement (NAT-SE) to alleviate the mismatch problem without target-domain data. Experiment results show that NAT-SE outperforms the best baseline Conv-TasNet in terms of PESQ and SSNR. The proposed disentangler module also improves

other encoder-decoder-like structure, such as WaveUnet. The subjective evaluation shows that the NAT-SE is significantly preferred over Conv-TasNet.

Chapter 4

Speech Enhancement with Multi-task Learning

This chapter[1] focuses on alleviating the mismatch problem caused by channel effect. Section 4.1 introduces the motivation and related studies. Section 4.2 combines the bandwidth extension module and the speech enhancement module for the condition that the high-frequency information is missing by channel effect. Section 4.3 reports the experimental setup and results. Section 4.4 concludes this chapter.

4.1 Motivation

In real-world life, the signals are sometimes not only distorted by various background noises but also missing the high-frequency information effected by transmission channels, such as high frequency (HF), very high frequency (VHF) and ultra high frequency (UHF). It is well known that speech signals with broader bandwidth provide higher perceptual quality and intelligibility, therefore, recovering the missing frequency information is important for speech enhancement task facing the mismatch problem caused by the channel effect.

Bandwidth extension is commonly utilized to recover the high-frequency information from narrowband signals, which is found useful in hearing aids design [73, 107], speech recognition [108–110] and speaker verification [111, 112].

[1]The study in this chapter has been published in [106]

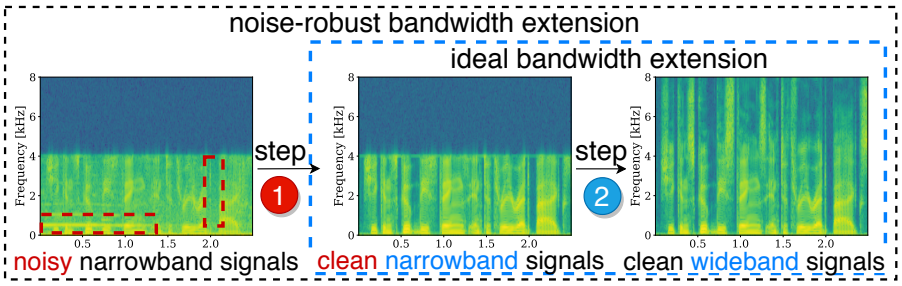


FIGURE 4.1: The work flow of speech enhancement and bandwidth extension tasks. In Step 1, the noisy narrowband signal is enhanced to remove noise. In Step 2, the enhanced narrowband signal is bandwidth-extended to generate the clean wideband signal.

Prior speech bandwidth extension methods, such as deep neural networks (DNN) [113, 114], fully convolutional network [115, 116], generative adversarial network (GAN) [117], and wavenet [118], mostly perform extension under ideal conditions with clean narrowband signals as inputs. This is called ideal bandwidth extension. However, in practice, when the signals are both distorted and partially-missing (i.e., the mismatch problem caused by channel effect), the only bandwidth extension module cannot handle this condition.

A typical way to address the mismatch problem caused by channel effect is to perform speech enhancement on the noisy narrowband signal first (Step 1), and ideal bandwidth extension next (Step 2), as illustrated in Figure 4.1. For example, there was a study to apply the iterative Vector Taylor Series (VTS) approximation algorithm [119] for feature enhancement, which is followed by a Gaussian mixture models or maximum a posterior models to reconstruct the wideband signals [120, 121].

With the advent of deep learning, recent studies suggest [47] an unified approach that combines speech enhancement and bandwidth extension (UEE) in a joint training neural network. As shown in Figure 4.2(a), the UEE approach firstly applies a bi-directional long-short-term-memory (BLSTM) layer as the speech enhancement module to map the noisy narrowband input to enhanced narrowband features. Then, another BLSTM layer is applied as the ideal bandwidth extension module [122] to recover the missing high-frequency information from the enhanced narrowband features. The speech enhancement and bandwidth extension module are first trained separately as the pre-training, which are then fine-tuned with a single mean square error (MSE) loss between the clean wideband ground-truth and

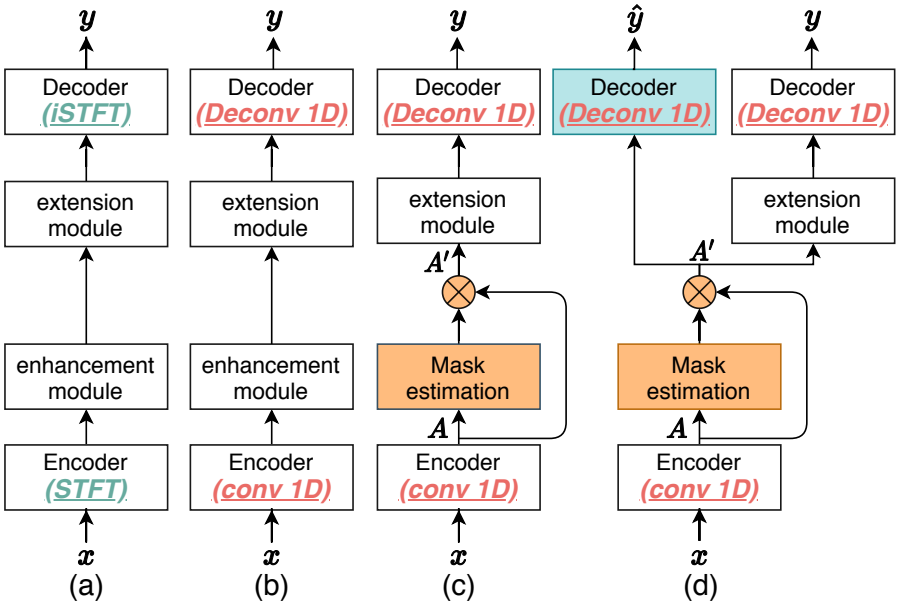


FIGURE 4.2: Block diagrams of (a) frequency-domain enhancement and extension, (b) time-domain enhancement and extension, (c) time-domain mask-based enhancement and extension (MBE), and (d) time-domain mask-based enhancement and extension with multi-task learning (MTL-MBE). $\otimes$ is an operator that refers to the element-wise multiplication.

enhanced-plus-extended output. Overall, the UEE approach is implemented with a two-stage training scheme, and it also faces phase estimation difficulty just like other frequency domain techniques.

In this chapter, we propose an end-to-end time-domain framework, which is achieved by jointly optimizing mask-based speech enhancement and ideal bandwidth extension modules with a multi-task learning (MTL-MBE). As a time-domain technique, the proposed method inherently avoids phase estimation issues. Specifically, the noisy narrowband signal is firstly encoded into acoustic features instead of the short time Fourier transform (STFT). The speech enhancement module takes the acoustic features to estimate a mask and obtains the enhanced narrowband features for subsequent bandwidth extension. Two speech decoders are trained to reconstruct the enhanced narrowband and enhanced-plus-extended features into time-domain signals, in a similar way like what inverse STFT (iSTFT) does. The network is optimized with a multi-task learning [123–125] over both narrowband and wideband signals.

4.2 Enhancement and extension

We now propose a time-domain masking and bandwidth extension modules with multi-task learning (MTL-MBE), which is illustrated in Figure 4.2 (d).

We first examine an enhancement and extension network in the time domain, which consists of a 1-D convolutional encoder to extract acoustic features from input speech, and a 1-D de-convolutional decoder to reconstruct waveforms from enhanced-plus-extended features, as shown in Figure 4.2(b). Such convolutional encoder-decoder-like structure is widely used in enhancement and separation tasks [1, 126]. The enhancement and extension are implemented as a pipeline of two similar regression, or mapping-based, neural networks. If trained jointly, their individual functions of the respective network are not clear. If trained separately, we face the same issue as other two-stage training schemes do.

4.2.1 Time-domain masking

To address the problem in the pipeline scheme of Figure 4.2(b), we propose a time-domain masking module to replace the mapping-based enhancement module, as shown in Figure 4.2(c), which has a unique architecture different from the extension module and is called MBE.

The time-domain masking aims to reduce the additive noise in noisy narrowband signals prior to extension. As shown in Figure 4.2(c), the input narrowband signal $x(t) \in \mathbb{R}^{1\times T}$ is encoded to a representation $A \in \mathbb{R}^{K\times M}$ by a 1-D convolutional layer. Such layer consists of $M(= 512)$ filters with a filter size of $L(= 16)$ samples and a stride of $L/2$ samples, which is followed by a rectified linear unit (ReLU) active function. Then, a time-domain masking W is estimated to suppress the additive noise in encoder representation A. It can be formulated as

$$A' = W \otimes A \tag{4.1}$$

where the estimated mask has the constraint $W \in [0, 1]$, $\otimes$ denotes element-wise multiplication, and A' is the enhanced representation output from the mask estimation module.

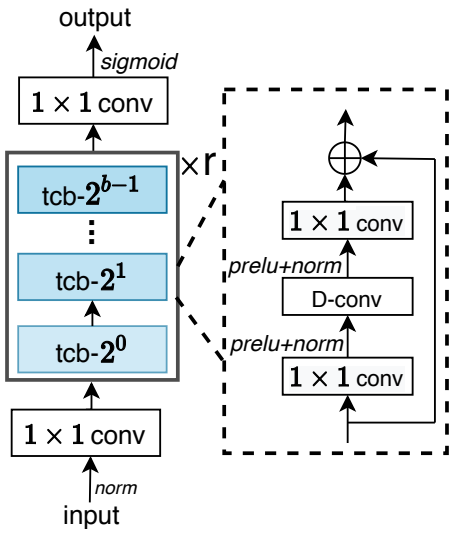


FIGURE 4.3: Block diagram of temporal convolutional network (TCN). "*tcb*-2^{b-1}" denotes a temporal convolutional block (TCB) with the dilation of 2^{b-1}, where b is the total number of the TCB. "D-conv" is the dilated convolutional layers stacked in several TCBs to exponentially increase the dilation factors. $\bigoplus$ is the residual connection.

The mask estimation module consists of a temporal convolutional network (TCN), which is illustrated in Figure 4.3. TCN is not the first time to be explored in speech enhancement. Prior work [127] utilized TCN as a regression module to map noisy input to clean signals, but their mapping-based framework is not suitable as an enhancement module here because it still suffers from the same problem as two-stage training schemes do. Therefore, we utilize TCN as a mask estimation module, which is a unique architecture different from the extension module.

As shown in Figure 4.3, the encoder representation A is firstly performed zero-normalization on channel dimension scaled by the learnable bias and gain variables [128]. Then, such representations are adjusted the channel numbers by a 1×1 convolutional layer with $N(= 128)$ filters. To learn the long-range temporal information of the speech with few number of parameters, we stack dilated convolutional layers in several temporal convolutional blocks (TCB) with exponentially increased dilation factor. As shown in dot box of Figure 4.3, two 1×1 convolutional layers and one dilated convolutional layer are applied in each TCN block with a parametric rectified linear unit (PReLU) [129] activation function and normalization operation. The first 1×1 CNN (with 512 filters and 1×1 kernel size) determines the input channels and the second 1×1 CNN (with 128 filters and 1×1 kernel size) adjusts the output channels from the dilated convolutional layer (with 512 filters and 1×3 kernel size). $b(= 8)$ TCBs are formed as a batch and each batch

is repeated for $r(=3)$ times in the TCN of mask estimation module. In each batch, we increase the dilation factors of the deptwise convolutions in the b TCBs as $[2^0, \ldots, 2^{b-1}]$. To keep the estimated mask W in a consistent dimension with the encoder representations A, one 1×1 CNN (with 512 filters and 1×1 kernel size) is applied with a sigmoid activation function for ensuring that the estimated mask W ranges within $[0, 1]$.

4.2.2 Multi-task learning

To provide cogent constraints for the enhancement module training, we further propose a multi-task loss for MBE as shown in Figure 4.2(d), that is designed for two training objectives: enhancement ("en") and extension ("ex"). It can be formulated as

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{ex}(y, z) + (1 - \lambda)\mathcal{L}_{en}(\hat{y}, \hat{z}) \tag{4.2}$$

where y denotes the enhanced-plus-extended signal, while z is its corresponding clean wideband signal as ground-truth target for training; similarly $\hat{y}$ denotes the enhanced narrowband signal, while $\hat{z}$ is its corresponding clean narrowband signal as ground-truth target. All signals are sampled at 16kHz. λ is a trainable weighting parameter to balance the two loss functions. $\mathcal{L}_{ex}$ and $\mathcal{L}_{en}$ loss functions are optimized via scale-invariant signal-to-distortion ratio (SI-SDR) [102, 130, 131].

As shown in Figure 4.2(d), two loss functions are applied at different places of the processing pipeline. For enhancement objective, the enhanced representation A' is reconstructed to form an enhanced narrowband signal $\hat{y}$ by a 1-D de-convolutional decoder, which is supervised by $\hat{z}$. For extension objective, A' is taken by the extension module to form an enhanced-plus-extended signal y, which is supervised by the clean wideband signals z. The proposed network in Figure 4.2(d) is referred to as multi-task learning for mask-based bandwidth extension, or MTL-MBE.

The extension module consists of a TCN as shown in Figure 4.3, which is similar to the mask estimation module, except that there is no element-wise multiplication $\otimes$, and we use ReLU as the activation function for the last 1×1 CNN instead of a sigmoid function.

4.3 Experiments and results

4.3.1 Database

We conduct evaluations on the public dataset by Valentini et al.[96], which is widely used for speech enhancement and bandwidth extension [39, 42, 75, 81, 132]. This dataset consists of 11,572 mono audio samples for training and 824 mono audio samples for testing. The speech is sampled at 16kHz. The training dataset has 40 noisy conditions (10 noise types $\times$ 4 signal-to-noise (SNR) values). The test dataset has 20 noise types that are different from the training set (5 new noise types $\times$ 4 new SNR values). The 2 speakers in the test dataset do not overlap the 28 speakers in the training dataset. We prepare both narrowband and wideband noisy data at 16kHz. We also prepare the clean wideband signals as ground-truth for the extension training and the clean narrowband signals as ground-truth for the enhancement training.

4.3.2 Experimental setup

4.3.2.1 Network configuration

During the training stage, the noisy narrowband waveforms were cut to 2-second long segments ($T = 32,000$ samples) for batch training. The Adam algorithm [133] is utilized to optimize the network. The learning rate is set to 0.001. We also adopt early stopping scheme when the loss increased on the development set for 20 epochs.

TABLE 4.1: PESQ, CSIG, CBAK, COVL, STOI and LSD performances of the proposed time-domain masking and multi-task loss.

Methods / Metrics	Single-loss		Multi-loss
	MBE w/o mask	**MBE**	**MTL-MBE**
PESQ	2.02	2.46	2.55
CSIG	2.13	2.52	2.64
CBAK	2.11	3.14	3.21
COVL	2.04	2.38	2.46
STOI	0.92	0.94	0.94
LSD	2.82	2.44	2.29

TABLE 4.2: A comparison of different techniques. "Designed conditions" refers to the conditions the method is designed for (clean or noisy). We perform all tests under noisy conditions. "#Paras" denotes the number of parameters of the model. "Feature type" denotes the types of narrowband inputs. "Spectrum" means that the approach is performed in frequency domain, while "waveform" means that time-domain signals are directly taken as inputs.

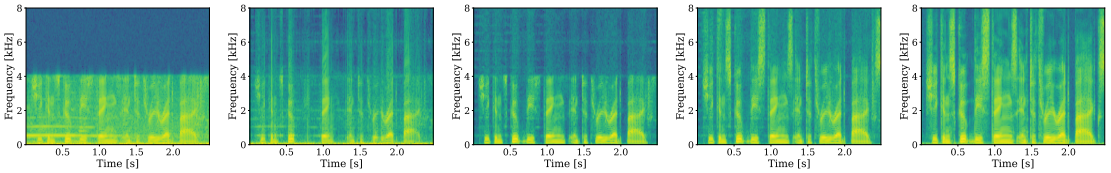
Designed conditions	Methods	#Paras	Feature type	PESQ	CSIG	CBAK	COVL	STOI	LSD
clean	LSM [113]	13.38M	spectrum	1.79	2.45	2.32	2.09	0.92	2.80
clean	DRCNN [116]	56.41M	waveform	1.74	1.18	1.97	1.38	0.92	2.97
noisy	UEE [47]	22.42M	spectrum	2.23	2.27	2.39	2.17	0.93	2.72
noisy	**MTL-MBE**	**6.82M**	waveform	**2.55**	**2.64**	**3.21**	**2.46**	**0.94**	**2.29**

4.3.2.2 Reference baselines

We implement three reference baselines. Two of them [113, 116] are for ideal bandwidth extension under noisy conditions. The other [47] is designed particularly for enhancement and extension.

- **LSM** [113]: a 3-layer network that predicted the missing high-frequency components from the low-frequency log-spectrum in frequency domain. The missing high-frequency phase was recovered by the imaged phase of low-frequency signals.
- **DRCNN** [116]: a fully convolutional encoder-decoder framework that mapped narrowband signals to wideband in the time domain. To increase the time dimensions during upscaling, subpixel shuffling layers were introduced in the upsampling blocks. The skip connections were utilized to speed up training.
- **UEE** [47]: a unified speech enhancement module and bandwidth extension module in one frequency-domain framework that recovered the high-frequency signals from noisy narrowband signals, as shown in Figure 4.2(a).

We use the following metrics to evaluate the results. PESQ [98] stands for perceptual evaluation of the speech quality, ranging from -0.5 to 4.5. Three objective metrics that approximate mean opinion scores (MOSs) [99]: CSIG, CBAK and COVL. They are designed for signal distortion evaluation, noise distortion evaluation, and overall quality evaluation, respectively. Short-time objective intelligibility (STOI) [56] reflects the improvement of speech intelligibility. Log-spectral distortion (LSD) [134] is to measure the distance between reconstructed and target spectrum. Except LSD, higher scores are better for all metrics.



(a) (b) (c) (d) (e)

FIGURE 4.4: The spectrograms of a sample (p232_005.wav) in the test set for (a) noisy-narrowband input, (b) the best baseline UEE, (c) enhanced narrowband result of MTL-MBE, (d) the enhanced-plus-extended result of MTL-MBE and (e) wideband signal (ground-truth).

4.3.3 Results

4.3.3.1 Effect of the proposed time-domain masking

We first investigate how the proposed time-domain masking contributes to the framework MBE in Figure4.2(c) by experimenting with and without (w/o) the time-domain mask. For fair comparison, the single loss is utilized in this experiment and the results are summarized in Table 4.1. We observe that the performances of MBE w/o time-domain masking decrease sharply because the noise issue is not addressed. Under the constraint of the single loss, the MBE achieves 21.8% and 13.5% relative improvements in terms of PESQ and LSD, compared with MBE w/o mask. The experiment also confirms the need to perform enhancement prior to bandwidth extension operation.

4.3.3.2 Effect of the proposed multi-task loss

We further investigate how the proposed multi-task learning contributes to the enhancement and extension. The comparative results of the MBE in Figure 4.2(c) and the MTL-MBE in Figure 4.2(d) are shown in Table 4.1. We observe that the performances are improved by utilizing the multi-task loss. Compared with the MBE, the MTL-MBE achieves 3.7% and 6.1% relative improvements in terms of PESQ and LSD. Such experiments show the performances of enhancement and extension can be further improved by providing constraints for the enhancement module.

4.3.3.3 Overall comparisons

Table 4.2 summarizes the comparison between the proposed MTL-MBE in Figure 4.2(d) and other baselines in terms of PESQ, CSIG, CBAK, COVL, STOI and LSD. "LSM" and "DPRNN" are designed for bandwidth extension under clean conditions but we evaluate them under noisy conditions in this experiment. Their results reveal the limitation when working under noisy conditions. We observe that the proposed MTL-MBE achieves the best performance. Comparing with the UEE method [47], MTL-MBE achieves 14.3% and 15.8% relative improvements in terms of PESQ and LSD. Meanwhile, the parameter size of MTL-MBE is 3 times smaller than that of UEE.

We extract one speech sample from the test set to illustrate the differences of recovered enhanced-plus-extended signal between the best baseline UEE and the proposed MTL-MBE, as shown in Figure 4.4. We observe that MTL-MBE (see Figure 4.4(d)) produces cleaner signal at low-frequency and richer high frequency content than UEE (see Figure 4.4(b)). The intermediate enhanced-narrowband magnitude spectrum is also shown in 4.4(c). We also observe that the enhanced-narrowband representations constrained by multi-task supervision provide well-presented features for subsequent extension operation.

4.3.3.4 Subjective evaluation

Since the UEE presents the best baseline performances in the objective evaluation in Table 4.2, we only conduct an A/B preference test between the UEE and the proposed MTL-MBE to evaluate the signal quality and intelligibility for listening. We randomly select 20 pairs of listening examples and invite 10 subjects to choose their preference according to the quality and intelligibility. The percentage of the preferences is shown in Figure 5.8. We observe that the listeners clearly preferred the proposed MTL-MBE with a preference score of 84% to the best baseline UEE with a preference score of 10%. Most subjects significantly preferred the reconstructed wideband signals by the MTL-MBE, because the MTL-MBE produces cleaner signals at low-frequency and richer high-frequency content. Some listening examples are available at Github[2].

[2] `https://nanahou.github.io/mtl-mbe/`

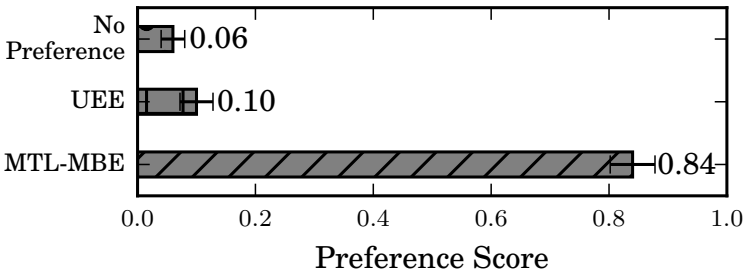


FIGURE 4.5: The result of A/B preference test for the recovered speech between the best baseline UEE and the proposed MTL-MBE.

4.4 Conclusion

In this chapter, we propose an end-to-end time-domain framework for enhancement and extension, that jointly optimizes a mask-based speech enhancement and an ideal bandwidth extension module with multi-task learning. The proposed framework avoids decomposing the signals into magnitude and phase spectra, therefore, requires no phase estimation. Experimental results show that the proposed method achieves 14.3% and 15.8% relative improvements over the best baseline in terms of perceptual evaluation of speech quality (PESQ) and log-spectral distortion (LSD), respectively. Furthermore, our method is 3 times more compact than the best baseline in terms of the number of parameters.

Chapter 5

Speech Enhancement with Hybrid Filterbanks Design

Single-channel time-domain speech enhancement has recently made great progress thanks to the learned filterbanks as used in Conv-TasNet. Such learned filterbanks can fully capture acoustic information from speech signals as well as avoid decomposing the speech signals into magnitude and phase spectra. The phase estimation can be omitted in this process. However, these approaches are usually trained fully relying on the training data, which are sensitive to varying testing scenarios.

In this chapter[1], we aim to address the mismatch problem caused by sensitive time-domain encoder/decoder. Specifically, Section 5.1 introduces the motivation of hybrid filter banks design and related studies. Section 5.2 presents the proposed speech enhancement with hybrid filterbanks design. Section 5.3 describes the experimental setup and results. Section 5.4 concludes this chapter.

5.1 Motivation

Humans have a remarkable ability to focus attention on a particular speech in noisy acoustic environments or even including other competing speech like the bubble noise. Speech enhancement algorithms mimics human's selective attention to mask out the noisy background environments and focus on the important

[1]The work in this chapter has been submitted in [135]

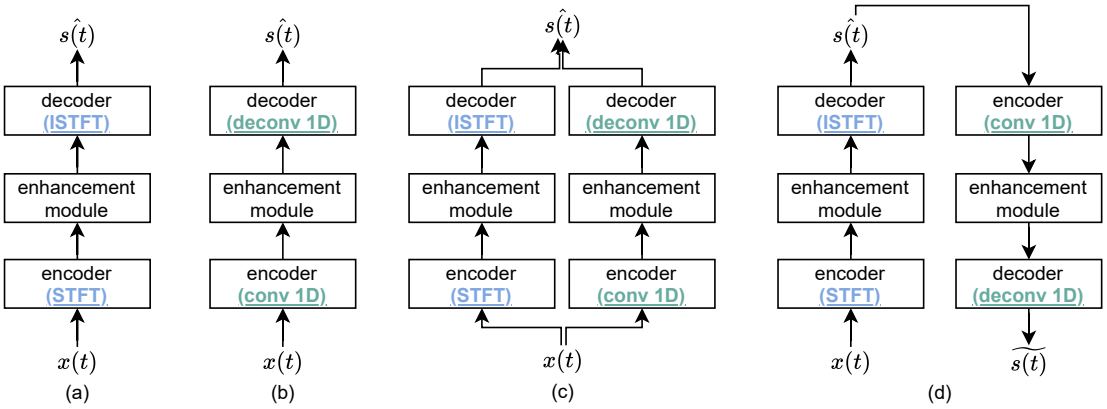


FIGURE 5.1: The block diagram of (a) frequency-domain speech enhancement network, (b) time-domain speech enhancement network, (c) two-branch multi-domain speech enhancement network, and (d) dual-path speech enhancement network (DTLN). $x(t)$ is the noisy speech and $\hat{s(t)}$ and $\widetilde{s(t)}$ denotes the predicted clean speech.

speech content. Such algorithms have served as a pre-processing module in many real-world applications, such as automatic speech recognition (ASR), speaker identification, and hearing aids design [8, 9].

For several decades, various statistical approaches were proposed to mimic the human's selective attention for masking out the noises, such as soft-decision noise suppress filter [12], minimum-mean-square-error (MMSE)-based spectral amplitude estimator [11], subspace algorithms [10], generalized gamma priors [13] and others [14–20]. Such algorithms attempted to estimate an optimal multiplicative masking through statistical inference to suppress noises.

With the advent of deep neural networks (DNNs), learning-based models were applied to predict clean speech from distorted inputs firstly in frequency domain, such as feed-forward networks [21–24], recurrent and long short-term memory (LSTM) networks [2, 25–27], as shown in Figure 5.1 (a). Some studies [1, 28] reveal that frequency-domain methods has several limitations. First, short-time Fourier transform (STFT) is a general signal transformation which is not proved as the most optimal feature extraction for speech enhancement. Second, accurate reconstruction of the phase for enhanced speech is a difficult problem, and the erroneous estimation of the phase leads to sub-optimal speech quality [29, 30]. Some post-processing methods are proposed to alleviate this problem by predicting phase information [31–35], but the final performance is still sub-optimal.

More recent end-to-end frameworks were proposed to avoid the phase manipulation issue by operating directly on the noisy waveform and eliminating an explicit STFT, such as fully convolutional networks (FCNs) [36–38], wave-U-Net architectures [39–43], Conv-Tasnet [1] and others [27, 36, 37, 44, 45], as shown in Figure 5.1 (b). Such time-domain end-to-end approaches can optimize the whole encoder-decoder like structure as well as omit the phase estimation. However, as the time-domain encoder and decoder are fully learned from the training data, they are sensitive to varying testing environments [16, 48, 49]. Recent studies also [50–52] observe that the time-domain speech separation/enhancement frameworks might not always yield good performances facing varying testing scenarios. The reasons caused such problem are in two-folds. First, the convolutional filters in the time-domain speech encoder fully rely on training data. They cannot always yield an good decomposition of the input speech compared with the fixed STFT algorithm facing various test data. Second, the end-to-end training strategy cannot guarantee that the speech encoder and decoder are well trained in this process, which also might cause the sensitivity of time-domain speech separation/enhancement frameworks (i.e., the mismatch problem caused by the sensitive time-domain encoder/decoder).

To alleviate this problem, prior work [3] attempts to directly combine the frequency-domain network and time-domain network as two branches to obtain the balanced results by averaging the two outputs respectively from the two branches, as shown in Figure 5.1 (c). Likewise, work [4] combines the frequency-domain network and time-domain network in a sequence to take the advantages of the two domains features, as shown in Figure 5.1 (d). However, the performance of such methods remains sub-optimal. The prior studies are the source of inspiration of this work.

In this chapter, we propose a two-step hybrid filterbanks design for time-domain speech enhancement (TSHFNet) to improve the robustness facing the varying testing environments. Firstly, we design the encoder and decoder consisting of hybrid filters, which can alleviate the problem that the speech encoder only consisting of conv-filters might not produce good features facing various test data. Secondly, we propose to optimize the whole framework using the two-step strategy, which helps to ensure that the speech encoder and decoder are well trained. Different from the previous work, this study makes the following contributions:

- We embed the conv-filters, param-filters and gamma-filters in the speech encoder and their inverse transformation in the speech decoder to guarantee

that learnable, semi-learnable, and non-learnable embedding coefficients can be learned. In this way, the proposed encoder and decoder are still considerable flexibility but rely less on the distribution of the data.

- We optimize the whole framework in two-stage. In the first step, we learn a transform and its inverse to a latent space where softmax-based masking is optimal. For the second step, we train an enhancement module that operates on the previously learned space. Two-step optimization ensures that every component is well-trained during the training stage.

5.2 Two-Step hybrid filterbanks design

A speech enhancement network generally consists of three network components: the speech encoder, the enhancement module and the speech decoder. The speech encoder encodes the noisy speech x(t) into feature representation via the STFT algorithm or convolutional filters. The enhancement module aims to produce the enhanced features usually in two ways: 1) directly mapping the noisy features to clean features; and 2) estimating masks to block the additive noise. We utilize the mask-based enhancement module in the proposed approach. Finally the speech decoder reconstructs the enhanced speech signals from enhanced feature via the iSTFT algorithm or de-convolutional filters.

Given that a noisy signal $x(t)$ of T samples is mixed by the clean voice s(t) and background noise $n(t)$, which is formulated as,

$$x(t) = s(t) + n(t), \quad t = 1, ..., T(1) \tag{5.1}$$

At the testing stage, given a noisy signal $x(t)$, we attempt to predict $\hat{s}(t)$ via the speech enhancement algorithm as close to $s(t)$ subject to an optimization criterion. We will introduce the proposed hybrid filterbanks design in section 5.2.1 and the proposed two-step optimization strategy in section 5.2.2 as follows.

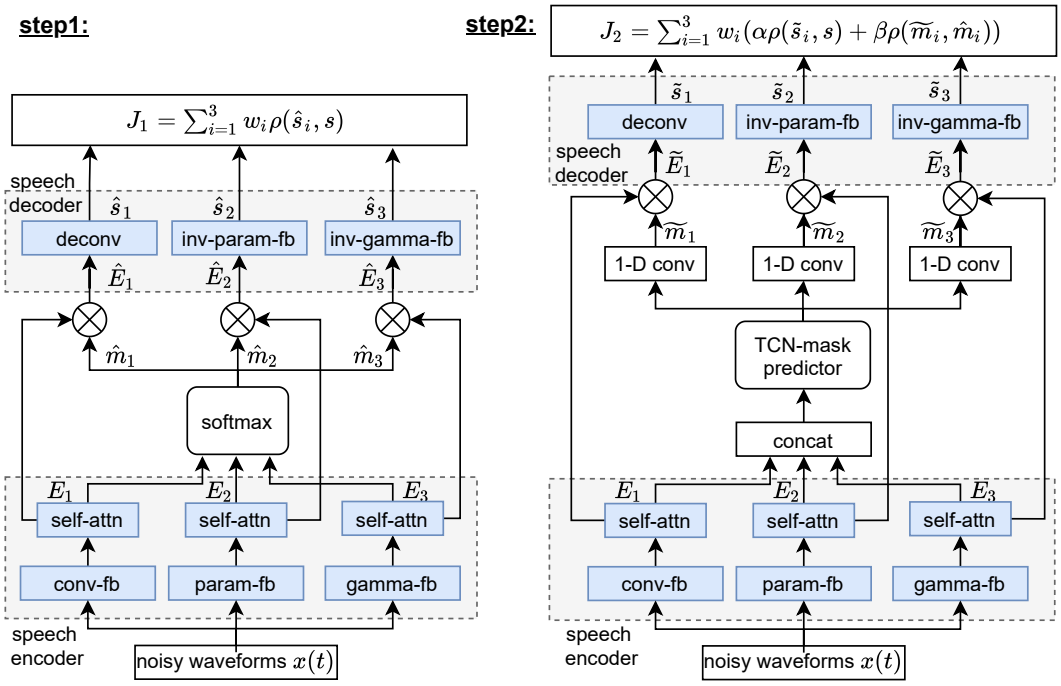


FIGURE 5.2: The block diagram of the proposed TSHFNet framework, that consists of two-step optimization. In each step, the proposed TSHFNET framework consists of a speech encoder (in gray), a mask predictor and a speech decoder (in gray). The speech encoder and speech decoder utilize the pretrained weights of those in step1. The mask predictor in step2 consists of the temporal convolutional network (TCN) rather than a simple softmax activate function in step1. The "conv-fb" and "deconv" are the 1-D convolutional filters and 1-D deconvolutional filters. The "param-fb" and "inv-param-fb" are the parameterized filterbank and inverse parameterized filterbank operations. The "gamma-fb" and "inv-gamma-fb" are the gammatone filterbanks and inverse gammatone filterbank operations. E_i is the latent representation of the noisy input produced by the speech encoder in step1. $\hat{m}_i$ and $\widetilde{m}_i$ are the predicted mask representations in step1 and step2. $\hat{s}_i$ and $\widetilde{s}_i$ are the reconstructed signals from the speech decoder in step1. s is the clean signal (ground-truth). $\otimes$ the element-wise multiplication. ρ denotes the enhancement loss of scale-invariant signal-to-distortion ratio (SI-SDR). α and β aim to balance the gradient loss between reconstructed signals and predicted masks. w_i is a trade-off parameter to balance the gradient loss of each filterbank design.

5.2.1 Hybrid filterbanks design

Time-domain speech enhancement frameworks usually utilize the convolutional filters as the speech encoder. They can avoid decomposing the speech signals into magnitude and phase spectra, therefore the phase estimation can be omitted in this process. As the convolutional-based speech encoder are fully learned from the

training data, they are also sensitive to produce stable embedding coefficients facing various test scenarios. To alleviate such problem, we propose a unified hybrid filterbanks based network, which consists of the speech encoder, the TCN-mask predictor and the speech decoder, as shown in "step2" of Figure 5.1.

5.2.1.1 Speech encoder

Different from the common convolutional speech encoder [1, 36, 136], the hybrid speech encoder consists of three groups of filters: conv-filters, param-filters and gamma-filters, which correspond to the learnable filters, semi-learnable filters and non-learnable filters.

The input noisy speech $x(t) \in R^{1\times T}$ is firstly encoded to embedding coefficients by conv-filters followed by the self-attention layer [137] as shown next:

$$E_1 = SA(W_{conv} \times x(t)), \quad t = 1, 2, \ldots, T \tag{5.2}$$

Where W_{conv} is the learnable weights of the 1-D convolutional filters. They are initialed randomly and learned from the training set. $SA(\cdot)$ denotes the self-attention layer. E_1 is the extracted embedding coefficients.

However, conv-filters fully rely on the training data compared with frequency-domain feature extraction algorithms like Fourier Transformer. Inspired by [138] in speaker recognition task, we propose the param-filters, which consists of sine and cosine functions and only high/low cut-off frequencies are required to learn from the data. We express the embedding coefficients extracted by param-filters as:

$$\begin{aligned} E_2 = SA(W_{param} \times x(t)), \quad t = 1, \ldots, T \\ W_{param} = w_{im} + j\mathcal{H}[w_{im}] \\ W_{im} = 2f_w sinc(2\pi f_w t) e^{-2j\pi f_c t} \\ f_w = f_2 - f_1 \\ f_c = (f_1 + f_2)/2 \end{aligned} \tag{5.3}$$

where f_1 and f_2 are the learnable high and low cut-off frequencies. The cut-off frequencies can be initialized randomly in the range [0, $f_s/2$], where f_s represents the sampling frequency of the input signal. $\mathcal{H}$ denotes the Hilbert transform, which

imparts a $-\pi/2$ phase shift to each positive frequency component to obtain shift-invariant representations like magnitude features of STFT. The param-filters can focus more in the lower part of the spectrum, where many crucial speech information are located. Moreover, they significantly reduce the numbers of parameters learned from the training data and increase the interpretability of the encoder and decoder.

Furthermore, we also embed the gamma-filters in the speech encoder, which can extract hand-crafted auditory features, are not affected by the training data, and has been proved to be quite effective in speech separation task. We get:

$$\begin{aligned} E_3 &= SA(W_{gamma} \times x(t)), \quad t = 1, \ldots, T \\ W_{gamma} &= \alpha t^{p-1} e^{-2\pi bt} cos(2\pi f_g t + \phi) \end{aligned} \tag{5.4}$$

Where f_g denotes the center frequency, ϕ is the phase shift, α is the amplitude, p is the filter order and b is the filter bandwidth parameter. The gammatone filters are deterministic, therefore, they can produce the stable embedding coefficients facing the varying test scenarios.

While various filters are proposed nowadays, we only study the three representative filters in this work, without loss of generality. The noisy speech $x(t)$ are first processed by the proposed hybrid speech encoder consisting of the conv-filters, param-filters and gamma-filters. Then, they are activated by the followed rectified linear unit (ReLU) activation function and the self-attention layer to obtain the speech embedding $E = [E_1, E_2, E_3] \in R^{K \times 3N}$. To concatenate the embeddings across channels, we utilize the same filter numbers N for each type of filters. Such concatenated embedding E are taken as inputs for the subsequent TCN-mask predictor.

5.2.1.2 The TCN-mask predictor

Broadbent's filter model [139] is one of the earliest theories of attention. Particular to psycho-acoustic experiments, physical properties such as color, loudness, and pitch are utilized to process the stimuli. Then certain stimuli passes through selective filters of listeners for further processing, which can be modelled by a mask, such as ideal binary mask (IBM)[140], ideal ratio mask (IRM)[141], ideal amplitude mask (IAM)[24], wiener-filter like mask (WFM)[142] and phase sensitive mask

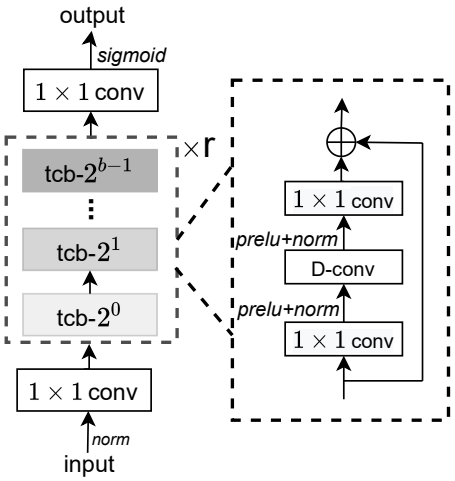


FIGURE 5.3: The block diagram of temporal convolutional network (TCN) based mask predictor. "$tcb - 2^{b-1}$" denotes the temporal convolutional block (TCB) with the dilation of 2^{b-1}, where b is the total number of the TCB. "D-conv" is the dilated convolutional layers stacked in several TCBs to exponentially increase the dilation factors. $\bigoplus$ is the residual connection.

(PSM)[142]. Likewise, we propose a TCN-mask predictor as shown in "step2" of Figure 5.2, which estimates receptive masks on the concatenated embeddings E. Then, we apply receptive masks $\widetilde{M_i}$ on embedding E_i and obtain enhanced coefficients $\widetilde{E_i}$ (named modulated responses [143] hereafter) for each filter design $(= 1, 2, 3)$:

$$\widetilde{E_i} = M_i \bigotimes E_i = f_M(E) \bigotimes E_i \tag{5.5}$$

where E is the concatenated embedding coefficients. $f_M(\cdot)$ are the functions representing the TCN-mask predictor. $\bigotimes$ presents element-wise multiplication.

The TCN-mask predictor consists of two 1-D CNN layers with 1×1 kernel size ("1×1 conv") and several stacked temporal convolutional blocks (TCBs) as shown in Figure 5.3. Specifically, we first conduct zero-normalization on the concatenated embeddings E along the channel dimension. We then apply the first "1×1 conv" layer on normalized embeddings and the sencond "1×1 conv" layer with O filters is utilized after TCBs to keep the dimensions of outputs same as inputs. To lean the long-term knowledge of inputs signals, we stack several TCBs between two "1×1 conv" layers with exponentially increased dilation factors like Conv-TasNet.

As shown in dot box of Figure 5.3, each TCB consists of two "1×1 conv" layers and one dilated depth-wise separable convolutional layer (i.e., "D-conv"). The first "1×1 conv" layer with P filters aims to transform input channels to P and the

second "1×1 conv" layer is applied to keep the dimensions of outputs same as inputs. The "D-conv" layer has P filters with a kernel size of $1 \times Q$ and a dilation factor of $2(B-1)$. We stack B TCBs in the TCN-mask predictor and repeat them for R times. Before applying predicted masks M_i on E_i, we adopt an another "1×1 conv" layer to maintain the dimensions of outputs from the last TCB same as inputs $E_i \in R^{K \times N}$. The estimated masks $M_i \in R^{K \times N}$ are then constrained in range of $[0, 1]$ via a sigmoid active function and the modulated responses $\widetilde{E_i} \in R^{K \times N}$ are finally estimated by Equation 5.5.

5.2.1.3 Speech decoder

Similar to the hybrid speech encoder, the proposed speech decoder also consists of three groups of filters: deconv-filters, inv-param-filters and inv-gamma-filters, which are inverse operation of conv-filters, param-filters and gammatone filters, respectively. As embedding coefficients for each filter design E_i lead to a modulated response $\widetilde{E_i}$, we reconstruct the hybrid modulated response $\widetilde{E_i}$ into time-domain signals $\widetilde{s_i}$ with different decoder functions.

5.2.1.4 Temporal convolutional network vs. self-attention layer

In this work, we adopt two popular mechanisms: temporal convolutional network (TCN) and self-attention layer to learn the long-range knowledge of the embedding coefficients. In the original Conv-TasNet [1], TCN is widely used in the mask-prediction module to learn longer temporal information with the depth-wise convolutional ("D-conv"), whose receptive field is "2^{b-1}". As the self-attention layer calculates the attention weights on the full embedding coefficients as shown in Equation 5.6, it could fully utilize the sequence-level dependency (i.e., global dependencies). Therefore, we adopt the self-attention layer in the speech encoder to further improve the performances of the TSHFNet.

$$SA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5.6}$$

Where Q, K, V are the linear transformations of the embedding coefficients outputted by the hybrid filters in the speech encoder and d_k is the dimension of the K.

5.2.2 Two-step optimization

The time-domain speech enhancement might be far from robust facing varying testing scenarios. One reason might be that the end-to-end training strategy cannot guarantee that the speech encoder and decoder are well trained in this process. To alleviate such problem, the prior work [144] proposed the two-stage network in the speech separation task, which separated the mixture speech at the first stage and then the estimated target speech are trained as close to the ground-truth inference at the second stage. Likewise, work [50] proposed an iterative training strategy to utilize the estimated target speech from the first network recursively as the inputs to the final separation network. Recent state-of-the-art results of Bert [145] in natural language processing tasks also show the benefits from pre-training the encoder transformation network in the first stage.

Inspired by the above studies, we propose to train the speech encoder and decoder ahead of the mask predictor to alleviate above problem, which is called two-step optimization strategy, depicted in Figure 5.2. We first train the hybrid speech encoder and decoder to obtain embedding coefficients $E = [E_1, E_2, E_3]$ for the noisy inputs. In step 2, we fix the hybrid speech encoder and decoder in step 1 and then train the TCN-mask predictor which is learned to estimate receptive masks $\widetilde{M} = [\hat{m}_1, \hat{m}_2, \hat{m}_3]$ for the embedding coefficients $E_i \in E$. Different from the prior pre-training strategy, we utilize the medium output of step 1 (i.e., predicted masks $\hat{m}_1$ in step 1) to supervise the training of the step 2.

5.2.2.1 Step 1: learning embedding coefficients

As a first step, we train the hybrid speech encoder $f_E(\cdot)$ to extract embedding coefficients for noisy inputs $x(t)$ and the speech decoder $f_D(\cdot)$ to reconstruct speech signals. Receptive masks $\hat{m}_i$ for the embedding coefficient E_i are estimated via a softmax function with noise segments as auxiliary inputs (across the dimension of the noisy speech and noise segments). Then, we obtain modulated responses $\hat{E}_i$ by:

$$\hat{E}_i = \hat{m}_i \bigotimes E_i \tag{5.7}$$

The decoder module $f_D(\cdot)$ is then trained to reconstruct speech signals $\hat{s}_i$ from modulated responses $\hat{E}_i$. At the training stage, we optimize the proposed network

via a SI-SDR loss [131], defined as J_1:

$$\begin{aligned} J_1 &= \sum_{i=1,2,3}^{3} w_i \rho(\hat{s}_i, s) \\ \rho(\hat{s}_i, s) &= 10 log_{10}(\frac{\| \frac{\langle \hat{s}_i, s \rangle}{\langle s, s \rangle} s \|^2}{\| \frac{\langle \hat{s}_i, s \rangle}{\langle s, s \rangle} s - \hat{s}_i \|^2}) \end{aligned} \tag{5.8}$$

where w_i are the weights. s and $\hat{s}_i$ denote the clean speech and the reconstructed speech, which are conducted zero-normalization before fed into SI-SDR loss function to maintain scale constancy. $\langle \cdot, \cdot \rangle$ denotes the inner product. The objective of this step is to find robust and stable embedding coefficients transformation, which facilitates speech enhancement through masking.

5.2.2.2 Step 2: training the TCN-mask predictor

Once the weights of the encoder and decoder modules are trained well in step 1, we fix their weights and train the TCN-mask predictor f_M as the step 2. Given the noisy input $x(t)$, the fixed speech encoder produces the embedding coefficients E. Then, f_M is trained to produce receptive masks $\widetilde{M}$ for the embedding coefficients E. Finally, the pre-trained decoder reconstruct the enhanced embedding coefficients $\widetilde{E}$ back into the time-domain signals $\widetilde{s}$.

Recent time-domain speech separation and enhancement approaches [1, 146] optimize the networks with the SI-SDR [131] as the loss function to calculate the differences between the clean signals and the reconstructed signals. For the training optimization of the step 2, we propose to apply SI-SDR loss function both on the time-domain signals and latent embedding coefficients, denoted as J_2:

$$J_2 = \sum_{i=1}^{3} w_i (\alpha \rho(\widetilde{s}_i, s) + \beta \rho(\widetilde{m_i}, \hat{m}_i)) \tag{5.9}$$

where α and β are the weights for the SI-SDR loss for time-domain signals and latent embedding coefficients, respectively. s and $\widetilde{s}_i$ denote the clean speech and the reconstructed speech. $\hat{m}_i$ and $\widetilde{m_i}$ denote the estimated receptive masks in the step 1 and step 2. The training procedures of the step 2 are similar with the step 1, while we also use the estimated masks $\hat{m}_i$ in step 1 as the parts of ground-truth labels to supervise the step 2 training.

Two advantages can benefit from the proposed two-step optimization strategy. Firstly, the TCN-mask can learn the intermediate knowledge in step 1, which reduces the difficulty of the training process. Secondly, measuring the distances between embedding coefficients also learns the local errors between each point.

5.3 Experiments and results

5.3.1 Database

To evaluate the performance of the proposed TSHFNet, we conducted experiments on three standard corpora: 1) DNS [147] challenge database of the INTERSPEECH 2020 version with a total number of 500 hours data, 2) DNS-20h database that includes 20 hours data randomly split from the DNS database for fast evaluation, and 3) VB database [148] with 1-hour data for evaluating the robustness of the proposed TSHFNet under various testing scenarios.

5.3.1.1 DNS corpus

DNS corpus [147] is the public dataset by Microsoft for the well-known deep noise suppression Challenge and we utilized the version of INTERSPEECH 2020 DNS challenge in this paper. The DNS corpus has two original collections: "clean", "noise", and "impulse-responses". The "clean" is approximate 500 hours containing 2150 speakers selected from Librivox [149], while the "noise" set includes over 180 hours of clips with 150 audio classes selected from Audioset [150], Freesound [151] and DEMAND database [54]. The DNS noisy speech corpus is created by adding clean speech and noise at various signal-to-noise-ratio (SNR) levels. We compute segmental SNR using segments in which both speech and noise are active. This is to avoid overshooting of amplitude levels in impulsive noise types such as door shutting, clatter, dog barking, etc. We synthesize 30s long clips by augmenting clean speech utterances and noise. The SNR levels are sampled from a uniform distribution between 0 and 40 dB. The mixed signal is then set to target Root Mean Square (RMS) level sampled from a uniform distribution between -15 dBFS and -35 dBFS [152, 153]. Finally, we obtained the 500-hours paired synthetic noisy data (i.e., noisy data, clean data, noise data).

The synthetic DNS consists of three subsets: training set, development set, and test set, where the development set is randomly selected from the training data in ratio of 10%. At the run-time inference, the DNS Challenge provides the publicly evaluation set, including two categories of synthetic clips. Similar to the prior work [1, 4, 74, 154, 155], we utilize the subset named "syn_noreverb" as the evaluation dataset. The evaluation set has 150 noisy clips with SNR levels distributed in between 0 dB to 20 dB and each clip is the 10-seconds segment. All the DNS speech signals are sampled at two sampling rates: 48kHz and 16kHz. We utilized the 16kHz data in this work.

5.3.1.2 DNS-20h corpus

As training the proposed TSHFNet with different configurations on the 500-hours corpus is resource-expensive and time-consuming, we randomly selected 20-hours data from DNS corpus for faster evaluation, that is called DNS-20h corpus. Likewise, the 10% data was selected from the DNS-20h corpus as the development set and the left data was utilized as the training set. At the run-time inference, we still utilize the "syn_noreverb" as the evaluation set which is same as the DNS corpus.

5.3.1.3 VB database

To evaluate the robustness of the proposed TSHFNet under unseen testing environments, we also utilize the evaluation set of the publicly dataset dataset, which is introduced in section 4.3.

5.3.2 Experimental setup

5.3.2.1 Configuration of speech encoder

Similar to prior work [1], the proposed hybrid speech encoder in Figure 5.2 extracts the embedding coefficients $E = [E_1, E_2, E_3]$ from the noisy inputs $x(t)$ by three groups of filters: conv-filters, param-filters and gamma-filters with equivalent filter number $N(= 512)$, followed by a ReLU activation function. To learn hybrid embedding coefficients, the three groups of filters had filter lengths of $L(= 16)$ with

$L/2(= 8)$ samples filter stride. As the filter length of the param-filters has to be odd, we set it as 17. The followed self-attention layer consists of three convolutional layers with 1×1 kernel size and the softmax activated function is utilized for obtaining the attention maps.

5.3.2.2 Configuration of TCN-mask predictor

The embedding coefficients E are first performed zero-normalization with learnable gain and bias parameters applied to E on the channel dimension. Then, the normalized features are fed in to the TCN-mask predictor in Figure 5.3. A 1×1 convolutional layer linearly transformed normalized features to representations with $O(= 128)$ channels, which determined the dimension of inputs. The number of input channels P equals 512 and the kernel size $1 \times Q$ of each depth-wise convolutional layer is set to 1×3. We stack $B(= 8)$ TCBs as a batch and repeated it for $R(= 3)$ times.

5.3.2.3 Configuration of speech decoder

The speech decoder in Figure 5.2 reconstructs the time-domain speech signal $(\widetilde{s_1}, \widetilde{s_2}, \widetilde{s_3})$ from the modulated responses $\widetilde{E_1}, \widetilde{E_2}, \widetilde{E_3}$. The deconv-filters, inv-param-filters and inv-gamma-filters in the speech decoder has the same configuration as those in the speech encoder, where the number of filters (N) for each group is equal to 512 and the filter lengths are turned to be 16 except for the inv-param-filters, which is 17.

5.3.2.4 Configuration of two-step optimization

In the first step, we maximize SI-SDR between the estimated signals and the clean signals with SI-SDR$(s, \hat{s})$. In the second step, we evaluate the SI-SDR on the estimated signals and the clean signals with SI-SDR$(s, \hat{s})$ as well as the predicted masks in step 1 and step2 with SI-SDR$(\hat{M}_i, \widetilde{M_i})$. We adopt Adam algorithm [133] to optimize the whole network with the initial learning rate of 0.001. If the loss doesn't decrease on the validation data for 3 epochs, we reduce it by half and if the loss doesn't decrease on the validation data for 10 epochs, we apply an early

stopping scheme. We cut the utterances into 1s segments for batch training, and set the minibatch size to 8.

5.3.2.5 Reference baselines

We select 6 systems that represent the recent advances in single-channel speech enhancement as the baselines, and report their results for benchmarking. The baseline systems are reported by the DNS challenges, which demonstrate state-of-the-art performances of speech enhancement techniques on DNS database.

- NSNet[154]: This work proposed a real-time speech enhancement approach based on a compact recurrent neural network trained with a simple mean-square-error (MSE) based speech distortion weighted loss function. To improve the robustness of their approach, the proposed MSE-based learning strategy separately controls over the importance of the speech distortion versus noise reduction, i.e., SNR information. The objective quality and intelligibility are also considered in this process.

- DTLN[4]: This work introduces a dual-signal transformation LSTM network for real-time speech enhancement. To improve the robustness of the approach under varying testing scenarios, the DTLN combines a short-time Fourier transform (STFT) and a learned analysis and synthesis basis in a stacked-network approach. Combining these two types of signal transformations enables the DTLN to robustly extract information from magnitude spectra and incorporate phase information from the learned feature basis.

- TCNN[127]: This work proposes a fully convolutional neural network for real-time speech enhancement in the time-domain. Similar to work [1], it is an encoder-decoder based architecture with an additional temporal covolutional module (TCM) inserted between the encoder and decoder. The TCM uses the casual and dilated convolutional convolutional layers to capture the current and previous frames, which improves the robustness of the whole architecture.

- Sub-band Model[155]: This work proposes an output-delayed sub-band LSTM framework to reduce noises in frequency domain. Specifically, each frequency bin combines context frequencies of noisy signals as inputs for the LSTM framework. The output is the corresponding frequency bin of the clean

speech. To maintain the small model size, the LSTM network is trained across all frequencies.

- MultiScale+ [74]: This work proposes a phase-aware sigmoid mask (PHM) to tackle a denoising and dereverberation problem with a single-stage framework. Specifically, to predict clean phases, they reuse estimated magnitudes and respect the triangle inequality in complex domain between mixture, source and the rest. To balance the reverberation part in outputs at testing stage, they propose two PHMs to deal with direct and reverberant source. Furthermore, to improve the robustness of the framework, they optimize the whole network with a time-domain loss function.

- Conv-TasNet[1]: This work proposes a fully-convolutional time-domain network for end-to-end speech separation, which is applied to the enhancement task later. Conv-TasNet utilizes 1-D convolutional encoder to extract representations from noisy inputs and the noise is reduced by predicting a mask for encoded representations. The masked representations are then reconstructed to waveforms by 1-D de-convolutional decoder. The masks are predicted by temporal convolutional networks (TCN) consisting of several 1-D dilated convolutional blocks, which could learn the long-range knowledge of inputs while maintaining a small model size.

5.3.2.6 Metric

We follow the same evaluation metrics in the speech enhancement literature [81, 82] for ease of comparison. PESQ [156] stands for perceptual evaluation of the speech quality, ranging from -0.5 to 4.5. Three objective metrics that approximate mean opinion scores (MOSs) [55]: CSIG, CBAK and COVL. They are designed for signal distortion evaluation, noise distortion evaluation, and overall quality evaluation, respectively. Signal-to-distortion ratio (SDR) is also conducted for measuring the speech quality. Short-time objective intelligibility (STOI) [56] denotes improvement of speech intelligibility. Higher scores are better for all metrics. Furthermore, we also conduct A/B preference test to show speech perceptual quality of proposed method.

TABLE 5.1: PESQ, CSIG, CBAK, COVL, SSNR(dB), SDR(dB) and STOI performances between the one-step optimization and the two-step optimization."#Paras" is the parameters of the network. S is the clean signal (ground-truth). $\hat{E}$ and $\hat{M}$ represent the enhanced embedding coefficients and the predicted mask in step1. α and β aim to balance the gradient loss between reconstructed signals and predicted masks. All the systems are trained on DNS-20h corpus and tested on the evaluation set of DNS corpus.

System	Training paradigms	#Paras	Target	(α, β)	PESQ	CSIG	CBAK	COVL	SDR
1	one-step optimization	3.41M	S	–	(2.48) 1.91	(4.03) 3.37	(3.58) 2.69	(3.28) 2.61	(16.08) 12.58
2	two-step optimization	3.44M	S	–	(2.60) 2.02	(4.12) 3.48	(3.67) 2.78	(3.39) 2.73	(17.42) 13.86
3		3.44M	$\hat{E}$	–	(2.64) 2.05	(4.16) 3.52	(3.71) 2.82	(3.43) 2.77	(17.51) 13.90
4		3.44M	$\hat{M}$	–	(2.65) 2.05	(4.17) 3.52	(3.71) 2.81	(3.44) 2.77	(17.55) 13.92
5		3.44M	(S, $\hat{E}$)	(0.5, 0.5)	(2.62) 2.03	(4.15) 3.50	(3.69) 2.79	(3.41) 2.75	(17.46) 13.88
6		3.44M	(S, $\hat{M}$)	(0.5, 0.5)	(2.64) 2.04	(4.15) 3.50	(3.69) 2.79	(3.40) 2.75	(17.48) 13.89
7		3.44M	(S, $\hat{M}$)	(0.1, 0.9)	(2.65) 2.04	(4.17) 3.51	(3.71) 2.81	(3.41) 2.75	(17.52) 13.90
8		3.44M	**(S, $\hat{M}$)**	**(0.3, 0.7)**	**(2.67) 2.06**	**(4.22) 3.55**	**(3.73) 2.86**	**(3.46) 2.79**	**(17.59) 13.94**

5.3.3 Results

5.3.3.1 One-step optimization vs. two-step optimization

We first compare between two training schemes, one-step optimization and two-step optimization on DNS-20h corpus. For the one-step optimization, we train the speech encoder, the TCN-mask estimator and the speech decoder together from the scratch as same as the work [1]. For the two-step optimization, we train the speech encoder, the softmax mask predictor and the speech decoder in the first step. Then, we fix the speech encoder and decoder and train the TCN-mask estimator in the second step as shown in Figure 5.2. For fair comparison, we adopt the same structures of the speech encoder, the TCN-mask estimator and the speech decoder in both systems. The hybrid filterbanks design are not adopted in this experiments.

We observe from Table 5.1, that the systems with the two-step optimization (Systems 2-8) consistently outperform the system with the one-step optimization (System 1), especially when the systems with the two-step optimization have roughly same number of parameters as the system with the one-step optimization. The results clearly show the advantage of the two-step optimization over the one-step optimization for time-domain encoder-decoder-like speech enhancement methods. We consider that the better performance is contributed to 1) the fully trained-well speech encoder and decoder, and 2) the latent mask representations in step 1 as part of labels supervising the learning of the TCN-mask estimator in the second step.

TABLE 5.2: PESQ, CSIG, CBAK, COVL, SSNR(dB), SDR(dB), and STOI performances between the single filterbank and the hybrid filterbanks."#Paras" presents the parameters of the network. The "conv_A" and "deconv" are the 1-D convolutional with self-attention layer and de-convolutional filterbank. The "param_A" and "inv-param" are the parameterized filterbank with slef-attention layer and inverse parameterized filterbank. The "gamma_A" and "inv-gamma" are the gammatone filterbanks with slef-attention layer and inverse gammatone filterbank. "RS" means the reconstructed signals. $\hat{s}_w = w_1\hat{s}_1 + w_2\hat{s}_2 + w_3\hat{s}_3$. s_i is the enhanced signals. w_i is a trade-off parameter to balance the gradient loss of each filterbank design. All the systems are trained on DNS-20h corpus and tested on the evaluation set of DNS corpus.

System	Single vs. Hybrid Fbs			(w_1, w_2, w_3)	RS	#Paras	PESQ	CSIG	CBAK	COVL	SDR
	Speech Encoder Fbs										
	Fb_1	Fb_2	Fb_3								
8	conv	–	–	(1.0, 0, 0)	s_1	3.44M	(2.67) 2.06	(4.22) 3.55	(3.73) 2.86	(3.46) 2.79	(17.59) 13.94
9	conv_A	–	–	(1.0, 0, 0)	s_1	5.01M	(2.71) 2.10	(4.28) 3.60	(3.78) 2.90	(3.52) 2.84	(17.75) 14.10
10	–	param	–	(0, 1.0, 0)	s_2	3.43M	(2.75) 2.13	(4.32)3.62	(3.78) 2.91	(3.55) 2.86	(17.85) 14.22
11	–	param_A	–	(0, 1.0, 0)	s_2	5.00M	(2.78) 2.15	(4.35) 3.66	(3.82) 2.94	(3.61) 2.91	(17.90) 14.29
12	–	–	gamma	(0, 0, 1.0)	s_3	3.43M	(2.73) 2.09	(4.25) 3.58	(3.73) 2.87	(3.50) 2.82	(17.75) 14.09
13	–	–	gamma_A	(0, 0, 1.0)	s_3	5.00M	(2.75) 2.12	(4.33) 3.62	(3.78) 2.92	()3.55 2.87	(17.80) 14.16
14	conv_A	param_A	–	(0.5, 0.5, 0)	s_1	6.64M	(2.83) 2.19	(3.38) 3.68	(3.89) 2.99	(3.65) 2.93	(18.02) 14.36
15	–	param_A	gamma_A	(0, 0.5, 0.5)	s_1	6.63M	(2.85) 2.21	(3.42) 3.70	(3.93) 3.03	(3.70) 2.97	(18.07) 14.40
16	conv_A	–	gamma_A	(0.5, 0, 0.5)	s_1	6.64M	(2.78) 2.15	(4.35) 3.65	(3.82) 2.94	(3.61) 2.90	(17.96) 14.31
17	conv_A	param_A	gamma_A	(0.4, 0.3, 0.3)	s_1	8.24M	(2.90) 2.25	(4.43) 3.72	(3.96) 3.05	(3.74) 3.01	(18.15) 14.49
18	conv_A	param_A	gamma_A	(0.4, 0.3, 0.3)	s_2	8.24M	(2.87) 2.23	(4.40) 3.70	(3.93) 3.02	(3.71) 2.99	(18.12) 14.44
19	conv_A	param_A	gamma_A	(0.4, 0.3, 0.3)	s_3	8.24M	(2.84) 2.19	(3.38) 3.68	(3.89) 2.98	(3.65) 2.93	(18.03) 14.36
20	conv_A	param_A	gamma_A	(0.4, 0.3, 0.3)	s_w	8.24M	(2.91) 2.27	(4.45) 3.74	(3.99) 3.07	(3.79) 3.05	(18.21) 14.53
21	**conv_A**	**param_A**	**gamma_A**	**(0.3, 0.5, 0.2)**	s_w	**8.24M**	**(2.95) 2.29**	**(4.48) 3.79**	**(4.04) 3.10**	**(3.82) 3.07**	**(18.26) 14.57**

5.3.3.2 Effect of learning targets on two-step optimization

We further analyse and summarize the performances by training with different learning targets on two-step optimization, as shown in System 2-8 of Table 5.1. Comparison between System 2 and System 3 shows that using the latent representations $\hat{E}$ of the first step as learning targets achieve better performance than using clean speech S as learning targets. Comparison between System 3 and System 4 shows that using the latent representations $\hat{E}$ of the first step as learning targets produce roughly same performances than using the predicted masks $\hat{M}$ of the first step as learning targets. If the two-step optimization has multiple outputs like predicted clean waveforms $\hat{S}$ and their corresponding predicted masks $\hat{M}$, we could optimize the two-step optimization framework with a weighted SI-SDR loss, as defined in Equation 5.8. With multiple learning targets: clean waveforms S and the predicted mask in first step $\hat{M}$, the best performances of PESQ, CSIG, CBAK, COVL and SDR are achieved at 2.06, 3.55, 2.86, 2.79 and 13.94 dB when the weights α and β in Equation 5.9 are turned to be 0.3 and 0.7. The results clearly shows the advantage of the SI-SDR loss optimized on both time-domain space and latent embedding coefficients space with two-step optimization.

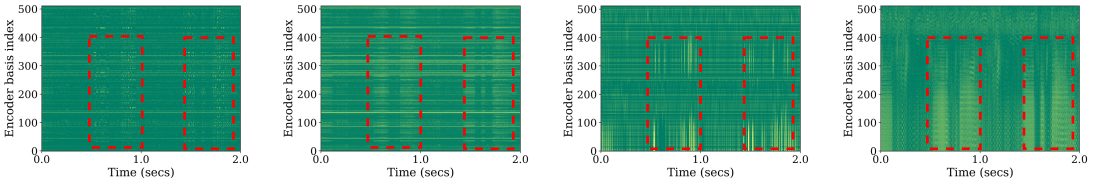


FIGURE 5.4: The illustrations of latent representations (sample: fileid_006.wav) in the test set for (a) the representation of conv-filter in one-step optimization (system 1), (b) the representation of conv-filter in two-step optimization (system 21), (c) the representation of param-filter in two-step optimization (system 21) and (d) the representation of gamma-filter in two-step optimization (system 21).

5.3.3.3 Single filterbank vs. hybrid filterbanks

Then, we validate the proposed hybrid filterbanks design. We continue to utilize the two-step optimization as the training strategy and integrate the various hybrid filterbanks in the speech encoder and decoder as shown in Figure 5.2. We observe from the Table 5.2, that the systems with the param-filters (system 10-11) outperform the systems with the conv-filters or gamma-filters (system 8-9 and 12-13), especially when the systems with the param-filters have roughly same number of parameters as the system with the conv-filters or gamma-filters. Compared with the systems with single types of filters (system 8, 10, 12), the corresponding systems with self-attention layer (system 9, 11, 13) can further improve the performances. Likewise, the systems with param-filters and gamma-filters (system 15) outperform the systems with the combinations of the other two (system 14, 16). The best performance is achieved by the system with three hybrid filters (system 21), which clearly shows the advantage of the hybrid filters to the robustness of the framework.

5.3.3.4 Effect of hybrid filters with various configurations

We further analyse and summarize the performances of the proposed TSHFNet with different weighted loss and different reconstructed signals, as shown in System (17-21) of Table 5.2. Comparison between System 17-20 shows that using the weighted reconstructed signals $s_w = w_1 * s_1 + w_2 * s_2 + w_3 * s_3$ performs better than the single reconstructed signals such as s_1, s_2 or s_3. The best performances are achieved with hybrid filterbanks design and weighted reconstructed signals when

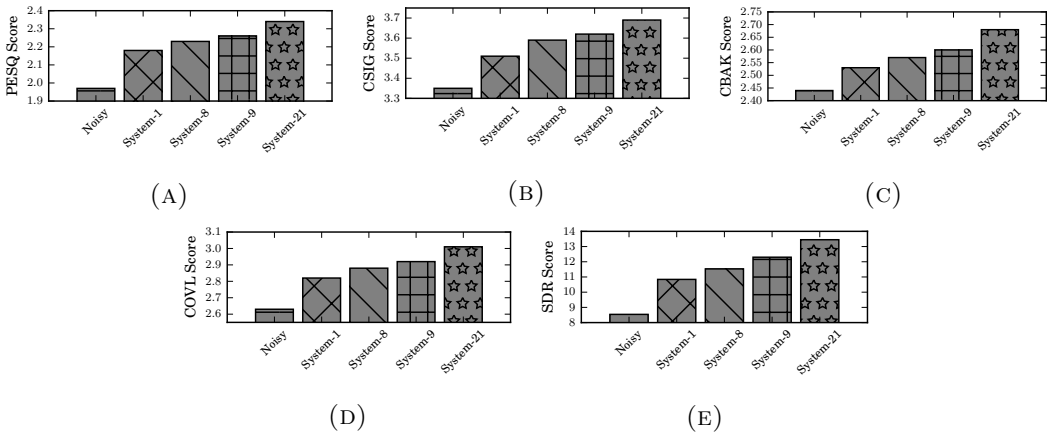


FIGURE 5.5: SDR(dB), PESQ, CSIG, CBAK, and COVL performances of the proposed TSHFNet approach under unseen testing scenarios. The system 1 denotes the Conv-TasNet baseline using the one-step optimization strategy. The system 8 can be seen as the Conv-TasNet baseline using the two-step optimization strategy. Compared with the system 8, the system 9 additionally utilizes the self-attention layer in speech encoder. The system 21 represents the best performances of the proposed TSHFNet. All the systems 1, 8, 9 and 21 are trained on DNS-20h corpus and evaluated on VB evaluation set.

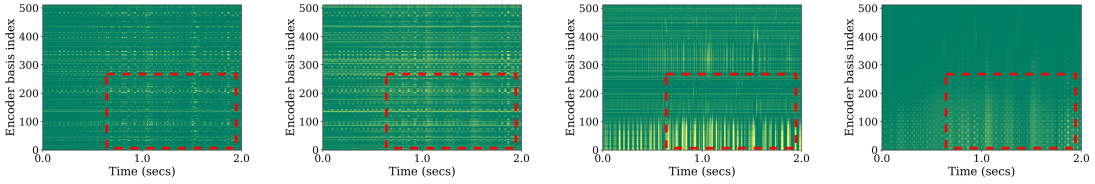


FIGURE 5.6: The illustrations of latent representations (sample: p232_005.wav) in the test set for (a) the representation of conv-filter in one-step optimization (system 1), (b) the representation of conv-filter in two-step optimization (system 21), (c) the representation of param-filter in two-step optimization (system 21) and (d) the representation of gamma-filter in two-step optimization (system 21).

the weights w_1, w_2 and w_3 in Equation 5.9 are tuned to be (0.3, 0.5, 0.2), which is aligned with the findings that the system 11 with param-filters also performs better the systems 9 and 13 under the same conditions.

5.3.3.5 Effect of hybrid filters under unseen testing scenarios

In this experiment, we would like to validate the robustness of the proposed TSHFNet (system 21) under unseen testing scenarios. We trained the system 21 with the DNS-20h dataset but evaluated it on the VB evaluation set, which exists the mismatch between the two datasets. The results are shown in Figure 5.5. We

TABLE 5.3: PESQ, SDR(dB), and STOI(%) performances of recent state-of-the-art techniques. "#Paras" is the parameters of the network. All methods are trained on DNS corpus and evaluated on "syn_noreverb" evaluation set.

Methods	#Paras	PESQ	SDR(dB)	STOI(%)
Noisy	–	1.58	9.09	91.52
NSNet [154]	5.1M	2.15	–	94.47
DTLN [4]	1.0M	2.34	16.54	94.76
TCNN [127]	–	2.34	16.86	–
Sub-band Model [155]	1.3M	2.37	–	94.24
MultiScale+ [74]	–	2.71	–	–
Conv-TasNet [1]	5.08M	2.73	17.12	94.83
TSHFNet	**8.24M**	**2.93**	**17.65**	**95.01**

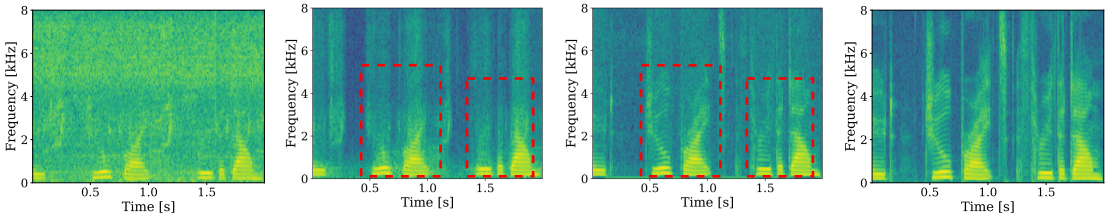


FIGURE 5.7: The spectrograms of a sample (fileid_006.wav) in the test set for (a) noisy input, (b) the best baseline (system 1), (c) enhanced result of TSFHNet (system 21) and (d) clean signal (ground-truth).

totally evaluate original noisy data and 4 systems (1, 8, 9, 21) in terms of PESQ, CSIG, CBAK, COVL and SDR on VB evaluation set. The results clearly show the advantages of the proposed TSHFNet even facing unseen testing scenarios.

To further show the robustness of the proposed TSHFNet, we illustrate the embedding coefficients of an example produced by the encoders of system 1, and 21 as shown in Figure 5.6. Comparison between the (a) of Figure 5.6 and (b) of Figure 5.6, the convolutional encoder with two-step optimization of system 21 still can extract more information than the convolutional encoder with the one-step optimization of system 1. Furthermore, the param-filters in the speech encoder of system 21 can extract the more clear information of noisy inputs, while the gamma-filters in the speech encoder of system 21 can accurately capture the low-frequency speech information and suppress the noise information.

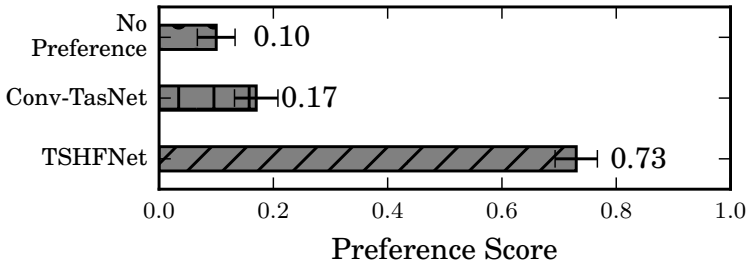


FIGURE 5.8: The result of A/B preference test for the enhanced speech between the proposed TSHFNet and the best baseline Conv-TasNet.

5.3.3.6 Benchmark against baselines

Table 5.3 summarizes the comparison between the proposed TSHFNet and other recent state-of-the-art techniques in terms of PESQ, SDR, and STOI. We observe that the proposed TSHFNet obtain the best performance. Comparing with the Conv-TasNet method, the TSHFNet achieves 7.3%, and 3.1% relative improvements in terms of PESQ and SDR.

Tor further show the contribution of the proposed TSHFNet, we illustrate the magnitude spectrum of an example as shown in Figure 5.7. We can see the proposed TSHFNet can reduce the approximate same noises and maintain the the richer low and high frequency information than the best baseline Conv-TasNet.

5.3.3.7 Subjective evaluation

As the Conv-TasNet produces the best baseline performance shown in Figure 5.3, we directly perform an A/B preference test between the proposed TSHFNet and Conv-TasNet method. We invited 10 listeners to choose their preference in 20 pairs of audio. We conclude the results in Figure 5.8. We observe that the listeners clearly preferred the proposed TSHFNet with a preference score of 73% to the best baseline Conv-TasNet with a preference score of 17%. Most subjects significantly preferred the enhanced speech signals by TSHFNet, because the audios sound more natural and have better quality.

5.4 Conclusion

In this chapter, we proposed a two-step hybrid filterbanks network (TSHFNet) to address the mismatch problem caused by the sensitive time-domain encoder/decoder. Specifically, we proposed to embed the conv-filters, param-filters and gamma-filters in the speech encoder to reduce the dependency of training data. Further more, we also propose to utilize the two-step training strategy to train a good encoder and decoder. Experimental results show that the proposed TSHFNet outperforms the best baseline Conv-TasNet in terms of PESQ, SDR and STOI.

Chapter 6

Conclusion and Future Work

Speech enhancement plays an important role in improving speech quality and intelligibility. In real-world situations, test data and testing environments are usually different from that of training data, which causes significantly degraded speech enhancement results. Many factors cause the mismatch problem in speech enhancement, but in this thesis, we focus only on unseen noises in testset, channel effect and sensitive time-domain encoder/decoder. To address mismatch problem, this thesis proposed three novel approaches to advance the performance of state-of-the-art speech enhancement techniques. Specifically, the three works will be summarized in Section 6.1. Section 6.2 will clarify the future work.

6.1 Conclusion

6.1.1 Speech enhancement with adversarial training

Speech denoising techniques aim to reduce excessive background noise in speech signals for better speech intelligibility and quality. For such machine-learning tasks, the training and testing data are usually assumed to have the same probability distribution. However, real-world situations often fail to meet this assumption, a consequence of unseen noises in the testset. As a result, speech enhancement performance may degrade significantly when faced with such unseen noises at runtime. In this chapter, we aim to alleviate the mismatch problem caused by unseen

noises under two real-world scenarios: with noisy target-domain data and without any target-domain data.

- If only noisy target-domain data is available, we propose a domain adversarial training technique for unsupervised domain transfer, that 1) overcomes domain mismatch, and 2) provides a solution to the scenario where we only have noisy speech data, and we don't have clean-noisy parallel data in the new domain. Specifically, our method includes two parts that are jointly trained, 1) an enhancement net to map noisy speech to clean speech by indirectly estimating a mask with a spectrum approximation loss, and 2) a domain predictor to distinguish between domains. Experiments suggest that our approach delivers voice quality comparable with other supervised learning techniques that require clean-noisy parallel data.

  Experiments are carried out using the CHiME-4 database and Valentini-Botinhao database. It is observed that our approach delivers voice quality comparable with other supervised learning techniques that require clean-noisy parallel data. Specifically, SE-DAT achieves relative improvements of 14.7% in terms of PESQ, as compared with the noisy speech baseline. Meanwhile, SE-DAT achieves an improvement of 6.6% PESQ relative to the no-adaptation baseline (SE-DAT-0). It indicates that adaptation via adversarial training plays an important role for better speech quality under conditions of mismatch. An A/B preference subjective listening test shows a 84% preference for the proposed SE-DAT compared with a 10% preference for the SE-DAT-0 baseline.

- If no target-domain data is available, we propose to learn noise-agnostic feature representations by disentanglement learning, which removes the unspecified noise factor, while keeping the specified factors of variation associated with clean speech. Specifically, a discriminator module is introduced to distinguish the type of noises, which is referred to as the disentangler. With the adversarial training strategy, a gradient reversal layer seeks to disentangle the noise factor and remove it from the feature representation. The network seeks to address the problem of unseen noise in the target domain without the need for any target domain data.

  Experiments are carried out using the Valentini-Botinhao database. We observe that the proposed NAT-SE obtained the best performances. Compared

with the best baseline (Conv-TasNet), the NAT-SE achieves 5.8% and 5.2% relative improvements in terms of PESQ and SSNR, respectively. In addition, we observe that listeners clearly preferred the proposed NAT-SE, with a preference score of 71% compared to the best baseline Conv-TasNet, with a preference score of 20%. Most subjects have a significant preference for the enhanced waveforms by NAT-SE, because there is less mismatch in the proposed NAT-SE.

6.1.2 Speech enhancement with multi-task training

In real-world situations, signals are sometimes not only distorted by various background noises but are also lacking the high-frequency information effected by transmission channels, such as high frequency (HF), very high frequency (VHF) and ultra high frequency (UHF). It is well known that speech signals with broader bandwidth provide higher perceptual quality and intelligibility, therefore, recovering the missing frequency information is important to speech enhancement when faced with the mismatch problem caused by the channel effect. To alleviate the mismatch problem caused by channel effect, the thesis further proposes an end-to-end time-domain framework for speech enhancement and bandwidth extension. Specifically, we jointly optimize mask-based speech enhancement and the ideal bandwidth extension module with multi-task learning. The proposed framework avoids decomposing the signals into magnitude and phase spectra, and therefore requires no phase estimation.

We conduct experiments on the Valentini-Botinhao database. Experimental results show that the proposed method achieves 14.3% and 15.8% relative improvements over the best baseline UEE in terms of PESQ and LSD, respectively. Furthermore, our method is 3 times more compact than the best baseline (UEE) in terms of the number of parameters. In addition, we also observe that the listeners clearly preferred the proposed MTL-MBE with a preference score of 84% to the best baseline UEE with a preference score of 10%. Most test subjects have a significant preference for wideband signals reconstructed by the MTL-MBE, because the MTL-MBE produces cleaner signals at low-frequency and richer high-frequency content.

6.1.3 Speech enhancement with hybrid filterbanks design

Single-channel time-domain speech enhancement has recently made great progress thanks to the learned filterbanks as used in Conv-TasNet. Such learned filterbanks can fully capture acoustic information from speech signals, as well as avoid decomposing the speech signals into magnitude and phase spectra. Phase estimation can be omitted in this process. However, these approaches are usually trained by fully relying on the training data, which is sensitive to varying testing situations. Therefore, this thesis further proposes a two-step hybrid filterbanks based network (TSHFNet) that can keep the advantages of time-domain approaches, while improving their robustness when faced with the mismatched problem caused by sensitive time-domain encoder/decoder. Specifically, TSHFNet consists of a hybrid speech encoder, a mask predictor, and a hybrid speech decoder. The hybrid speech encoder transforms noisy speech into three groups of embedding coefficients via learnable filters, semi-learnable filters and non-learnable filters in the encoder. The mask predictor takes the hybrid embedding coefficients as inputs and estimates masks for them. Finally, the hybrid speech decoder reconstructs the cleaned speech signals from the enhanced hybrid embedding coefficients. In the training stage, we propose a two-step optimization strategy for the good encoder and decoder.

We conduct the experiments on the DNS challenge database. Experimental results show that the proposed TSHFNet achieves 7.3% and 3.1% relative improvements over the best baseline Conv-TasNet on the DNS corpus in terms of perceptual evaluation speech quality (PESQ) and signal-to-distortion ratio (SDR). The ablation study indicates that the proposed TSHFNet can produce more robust embedding coefficients on matched/unmatched testing scenarios. In addition, we observe that the listeners clearly preferred the proposed TSHFNet with a preference score of 73% to the best baseline Conv-TasNet with a preference score of 17%. Most test subjects have a significant preference for the enhanced speech signals by TSHFNet, because the signals sound more natural and have better quality.

6.2 Future work

Speech enhancement is a potential technique to enable real-world applications work well in noisy environments. We plan to extend the current works in the following

three research directions.

Mismatch problem caused by other factors: This thesis has shown that speech enhancement technologies, via adversarial training, can achieve significant improvement over the previous baseline studies. However, these works only focus on the mismatch problem caused by unseen noise. In real-world situations, there are many other factors that can cause the training-testing mismatch, such as, different accents and various languages. Therefore, the speech enhancement method with adversarial training can be improved by learning more general acoustic features, which can disentangle the accent information or language information in the learned features.

Bandwidth extension pre-processing: When we simulate the signals corrupted by channel effect, we usually apply a low-pass filter to the full-band signals to block high-frequency information. This causes the dividing line between low-frequency and high-frequency signals to be extremely tidy. However, in real-world applications, there is no hard line between low-frequency signals and high-frequency signals, which also introduces the problem of mismatch between training and testing. Therefore, the speech enhancement method with multi-task training can be improved using different low-pass filters to pre-process the training data, which can leverage the mismatch problem between the training stage and the inference stage in real-world situations.

Real-time speech enhancement: The proposed speech enhancement technologies perform off-line processing, but real-world applications require processing in real-time. For example, a delay of 3ms is noticeable in real-time applications and delays time of over 10ms are unacceptable. Therefore, the speech technologies proposed in this thesis can be further improved by reducing 1) the depth of layers and 2) the operational complexity of the framework, to satisfy the requirements of real-time applications .

List of Author's Publications

Conference

- Chen Chen, **Nana Hou**, Hu Yuchen and Eng Siong Chng, "Noise-robust Speech Recognition with 10 Minutes Unparalleled In-domain Data", *Submitted to IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022).*

- Hu Yuchen, **Nana Hou**, Chen Chen and Eng Siong Chng, "Interactive Feature Fusion for End-to-End Noise-Robust Speech Recognition", *Submitted to IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022).*

- Huang Yizheng, **Nana Hou**, and Nancy F. Chen, "Progressive Continual Learning for Spoken Keyword Spotting", *Submitted to IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022).*

- **Nana Hou**, Chenglin Xu, Eng Siong Chng, and Haizhou Li, "Learning Disentangled Feature Representations for Speech Enhancement via Adversarial Training", *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021).*

- Duo Ma, **Nana Hou**, Van Tung Pham, Haihua Xu, and Eng Siong Chng, "Multitask-Based Joint Learning Approach to Robust ASR For Radio Communication Speech", *in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2021).*

- Chen Chen, **Nana Hou**, Duo Ma, and Eng Siong Chng, "Time Domain Speech Enhancement with Attentive Multi-scale Approach", *in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2021).*

- **Nana Hou**, Chenglin Xu, Joey Tianyi Zhou, Van Tung Pham, Eng Siong Chng, and Haizhou Li, "Speaker and Phoneme-Aware Speech Bandwidth Extension with Residual Dual-Path Network", *in INTERSPEECH 2020.*

- **Nana Hou**, Chenglin Xu, Joey Tianyi Zhou, Eng Siong Chng, and Haizhou Li, "Multi-task Learning for End-to-end Noise-robust Bandwidth Extension", *in INTERSPEECH 2020.*

- Hao Xiang, Chenglin Xu, **Nana Hou**, Lei Xie, Eng Siong Chng, and Haizhou Li, "Time-Domain Neural Network Approach for Speech Bandwidth Extension", *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020).*

- **Nana Hou**, Chenglin Xu, Eng Siong Chng, and Haizhou Li, "Domain Adversarial Training for Speech Enhancement", *in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2019).*

- **Nana Hou**, Xiaohai Tian, Eng Siong Chng, Bin Ma, and Haizhou Li, "Improving Air Traffic Control Speech Intelligibility by Reducing Speaking Rate Effectively", *in International Conference on Asian Language Processing (IALP 2017).*

Journal

- **Nana Hou**, Eng Siong Chng, and Haizhou Li, "Two Step Training for Speech Enhancement with Multi-filterbank Design", *will be submitted to IEEE ACM Transactions on Audio, Speech and Language Processing, TASLP 2021.*

Bibliography

[1] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019. xiv, xix, xxiii, 1, 2, 5, 21, 23, 24, 26, 29, 30, 45, 46, 50, 56, 66, 67, 70, 73, 75, 77, 79, 80, 81, 85

[2] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International conference on latent variable analysis and signal separation*, pages 91–99. Springer, 2015. xix, 1, 21, 22, 23, 24, 66

[3] Jang-Hyun Kim, Jaejun Yoo, Sanghyuk Chun, Adrian Kim, and Jung-Woo Ha. Multi-domain processing via hybrid denoising networks for speech enhancement. *arXiv preprint arXiv:1812.08914*, 2018. xix, xx, 2, 3, 22, 23, 29, 31, 67

[4] Nils L Westhausen and Bernd T Meyer. Dual-signal transformation lstm network for real-time noise suppression. *arXiv preprint arXiv:2005.07551*, 2020. xix, 2, 3, 22, 23, 31, 67, 77, 79, 85

[5] Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 4:2, 2019. xix, 23, 24

[6] Kazuyuki Hara, Daisuke Saito, and Hayaru Shouno. Analysis of function of rectified linear unit used in deep learning. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2015. xix, 23, 25

[7] 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. *arXiv preprint arXiv:2008.00264*, 2020. xix, 22, 27

[8] Dong Yu, Li Deng, Jasha Droppo, Jian Wu, Yifan Gong, and Alex Acero. A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4041–4044. IEEE, 2008. 1, 7, 66

[9] Li-Ping Yang and Qian-Jie Fu. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. *The Journal of the Acoustical Society of America*, 117(3):1001–1004, 2005. 1, 7, 66

[10] Yariv Ephraim and Harry L Van Trees. A signal subspace approach for speech enhancement. *IEEE Transactions on speech and audio processing*, 3 (4):251–266, 1995. 1, 66

[11] 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*, 32(6):1109–1121, 1984. 1, 15, 16, 18, 19, 20, 66

[12] Robert McAulay and Marilyn Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(2):137–145, 1980. 1, 20, 66

[13] Jan S Erkelens, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1741–1752, 2007. 1, 66

[14] Yariv Ephraim and David Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE transactions on acoustics, speech, and signal processing*, 33(2):443–445, 1985. 1, 15, 66

[15] Israel Cohen. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *IEEE Signal processing letters*, 9(4): 113–116, 2002.

[16] Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura. Generalized gamma modeling of speech and its online estimation for speech enhancement. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 4, pages iv–181. IEEE, 2005. 5, 21, 31, 67

[17] Bengt J Borgström and Abeer Alwan. Log-spectral amplitude estimation with generalized gamma distributions for speech enhancement. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4756–4759. IEEE, 2011.

[18] Philipos C Loizou. Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. *IEEE Transactions on Speech and Audio Processing*, 13(5):857–869, 2005.

[19] Bin Chen and Philipos C Loizou. A laplacian-based mmse estimator for speech enhancement. *Speech communication*, 49(2):134–143, 2007.

[20] Eric Plourde and Benoît Champagne. Auditory-based spectral amplitude estimators for speech enhancement. *IEEE transactions on audio, speech, and language processing*, 16(8):1614–1623, 2008. 1, 66

[21] 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. 1, 21, 66

[22] 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*, 23(1):7–19, 2014.

[23] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[24] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 22(12):1849–1858, 2014. 1, 21, 66, 71

[25] Felix Weninger, John R Hershey, Jonathan Le Roux, and Björn Schuller. Discriminatively trained recurrent neural networks for single-channel speech separation. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 577–581. IEEE, 2014. 1, 21, 66

[26] Bengt J Borgström, Michael S Brandstein, and Robert B Dunn. Improving statistical model-based speech enhancement with deep neural networks. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 471–475. IEEE, 2018.

[27] Ziyue Zhao, Samy Elshamy, and Tim Fingscheidt. A perceptual weighting filter loss for dnn training in speech enhancement. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 229–233. IEEE, 2019. 1, 2, 21, 66, 67

[28] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. The importance of phase in speech enhancement. *speech communication*, 53(4):465–494, 2011. 1, 21, 30, 66

[29] Benjamin J Shannon and Kuldip K Paliwal. Role of phase estimation in speech enhancement. In *Ninth International Conference on Spoken Language Processing*, 2006. 2, 21, 30, 66

[30] Pejman Mowlaee and Josef Kulmer. Phase estimation in single-channel speech enhancement: Limits-potential. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(8):1283–1294, 2015. 2, 21, 30, 66

[31] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama. Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency. In *Proc. DAFx*, volume 10, pages 397–403, 2010. 2, 21, 30, 66

[32] Nicolas Sturmel, Laurent Daudet, et al. Signal reconstruction from stft magnitude: A state of the art. In *International conference on digital audio effects (DAFx)*, pages 375–386, 2011.

[33] Kehuang Li, Bo Wu, and Chin-Hui Lee. An iterative phase recovery framework with phase mask for spectral mapping with an application to speech enhancement. In *INTERSPEECH*, pages 3773–3777, 2016.

[34] Naijun Zheng and Xiao-Lei Zhang. Phase-aware speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):63–76, 2018.

[35] Kohei Yatabe, Yoshiki Masuyama, and Yasuhiro Oikawa. Rectified linear unit can assist griffin-lim phase recovery. In *2018 16th international workshop on acoustic signal enhancement (IWAENC)*, pages 555–559. IEEE, 2018. 2, 21, 30, 66

[36] 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*, 26(9):1570–1584, 2018. 2, 21, 67, 70

[37] Ashutosh Pandey and DeLiang Wang. A new framework for supervised speech enhancement in the time domain. In *Interspeech*, pages 1136–1140, 2018. 2, 21, 67

[38] Francois G Germain, Qifeng Chen, and Vladlen Koltun. Speech denoising with deep feature losses. *arXiv preprint arXiv:1806.10522*, 2018. 2, 21, 67

[39] Craig Macartney and Tillman Weyde. Improved speech enhancement with the wave-u-net. *arXiv preprint arXiv:1811.11307*, 2018. 2, 21, 39, 43, 50, 59, 67

[40] Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy. Attention wave-u-net for speech enhancement. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 249–253. IEEE, 2019.

[41] 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*, 28: 825–838, 2020.

[42] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017. 39, 50, 59

[43] Fan Yang, Ziteng Wang, Junfeng Li, Risheng Xia, and Yonghong Yan. Improving generative adversarial networks for speech enhancement through regularization of latent representations. *Speech Communication*, 118:1–9, 2020. 2, 21, 67

[44] Yan Zhao, Buye Xu, Ritwik Giri, and Tao Zhang. Perceptually guided speech enhancement using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5074–5078. IEEE, 2018. 2, 21, 67

[45] Szu-Wei Fu, Chien-Feng Liao, and Yu Tsao. Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality. *IEEE Signal Processing Letters*, 27:26–30, 2019. 2, 21, 67

[46] Santiago Pascual, Maruchan Park, Joan Serrà, Antonio Bonafonte, and Kang-Hun Ahn. Language and noise transfer in speech enhancement generative adversarial network. *arXiv preprint arXiv:1712.06340*, 2017. 2, 34, 43

[47] Bin Liu, Jianhua Tao, and Yibin Zheng. A novel unified framework for speech enhancement and bandwidth extension based on jointly trained neural networks. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 11–15. IEEE, 2018. 2, 3, 54, 60, 62

[48] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160*, 2019. 5, 21, 31, 67

[49] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*, 2020. 5, 21, 31, 67

[50] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey. Universal sound separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 175–179. IEEE, 2019. 5, 22, 31, 67, 74

[51] Andong Li, Wenzhe Liu, Chengshi Zheng, Cunhang Fan, and Xiaodong Li. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1829–1843, 2021.

[52] Yuzhou Liu and DeLiang Wang. Divide and conquer: A deep casa approach to talker-independent monaural speaker separation. *IEEE/ACM Transactions on audio, speech, and language processing*, 27(12):2092–2102, 2019. 5, 22, 31, 67

[53] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE, 2013. 10

[54] 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. 10, 40, 76

[55] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007. 13, 80

[56] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010. 13, 60, 80

[57] Arye Nehorai and Boaz Porat. Adaptive comb filtering for harmonic signal enhancement. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1124–1138, 1986. 15

[58] K Paliwal and Anjan Basu. A speech enhancement method based on kalman filtering. In *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 177–180. IEEE, 1987. 15

[59] Xuemin Shen and Li Deng. A dynamic system approach to speech enhancement using the h/sub/spl infin//filtering algorithm. *IEEE Transactions on Speech and Audio Processing*, 7(4):391–399, 1999. 15

[60] Hossein Sameti, Hamid Sheikhzadeh, Li Deng, and Robert L Brennan. Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise. *IEEE Transactions on Speech and Audio processing*, 6(5):445–455, 1998. 15

[61] Yann Soon and Soo Ngee Koh. Speech enhancement using 2-d fourier transform. *IEEE Transactions on speech and audio processing*, 11(6):717–724, 2003. 15

[62] Yann Soon, Soo Ngee Koh, and Chai Kiat Yeo. Noisy speech enhancement using discrete cosine transform. *Speech communication*, 24(3):249–257, 1998. 15

[63] M Richards. Helium speech enhancement using the short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(6): 841–853, 1982. 15

[64] Navneet Upadhyay and Abhijit Karmakar. Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. *Procedia Computer Science*, 54:574–584, 2015. 15, 16

[65] Yariv Ephraim. A minimum mean square error approach for speech enhancement. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 829–832. IEEE, 1990. 15

[66] TV Sreenivas and Pradeep Kirnapure. Codebook constrained wiener filtering for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 4(5):383–389, 1996. 15, 16

[67] M Abd El-Fattah, Moawad Ibrahim Dessouky, Salah Diab, and Fathi Abd El-Samie. Speech enhancement using an adaptive wiener filtering approach. *Progress In Electromagnetics Research M*, 4:167–184, 2008. 15

[68] Thomas Lotter and Peter Vary. Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Advances in Signal Processing*, 2005(7):1–17, 2005. 15

[69] Yariv Ephraim. A bayesian estimation approach for speech enhancement using hidden markov models. *IEEE Transactions on Signal Processing*, 40 (4):725–735, 1992. 16

[70] Peter SK Hansen. *Signal subspace methods for speech enhancement*. PhD thesis, Citeseer, 1997. 16

[71] Nathalie Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on speech and audio processing*, 7(2):126–137, 1999. 16

[72] Hanjun Liu, Qin Zhao, Mingxi Wan, and Supin Wang. Enhancement of electrolarynx speech based on auditory masking. *IEEE Transactions on Biomedical Engineering*, 53(5):865–874, 2006. 16

[73] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2007. 19, 41, 53

[74] Hyeong-Seok Choi, Hoon Heo, Jie Hwan Lee, and Kyogu Lee. Phase-aware single-stage speech denoising and dereverberation with u-net. *arXiv preprint arXiv:2006.00687*, 2020. 21, 77, 80, 85

[75] Neil Shah, Hemant A Patil, and Meet H Soni. Time-frequency mask-based speech enhancement using convolutional generative adversarial network. In *Proceedings, APSIPA Annual Summit and Conference*, volume 2018, pages 12–15, 2018. 21, 43, 50, 59

[76] Shubo Lv, Yanxin Hu, Shimin Zhang, and Lei Xie. Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement. *arXiv preprint arXiv:2106.08672*, 2021. 22

[77] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. Phase-aware speech enhancement with deep complex u-net. In *International Conference on Learning Representations*, 2018. 22, 29

[78] Ke Tan and DeLiang Wang. Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6865–6869. IEEE, 2019. 27

[79] Yuxiang Kong, Jian Wu, Quandong Wang, Peng Gao, Weiji Zhuang, Yujun Wang, and Lei Xie. Multi-channel automatic speech recognition using deep complex unet. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 104–110. IEEE, 2021. 29

[80] Xuchen Song, Qiuqiang Kong, Xingjian Du, and Yuxuan Wang. Catnet: music source separation system with mix-audio augmentation. *arXiv preprint arXiv:2102.09966*, 2021. 29

[81] Nana Hou, Chenglin Xu, Eng Siong Chng, and Haizhou Li. Domain adversarial training for speech enhancement. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 667–672. IEEE, 2019. 33, 34, 59, 80

[82] Nana Hou, Chenglin Xu, Eng Siong Chng, and Haizhou Li. Learning disentangled feature representations for speech enhancement via adversarial training. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 666–670. IEEE, 2021. 33, 80

[83] Chien-Feng Liao, Yu Tsao, Hung-Yi Lee, and Hsin-Min Wang. Noise adaptive speech enhancement using domain adversarial training. *arXiv preprint arXiv:1807.07501*, 2018. 34, 35

[84] Qing Wang, Wei Rao, Sining Sun, Lei Xie, Eng Siong Chng, and Haizhou Li. Unsupervised domain adaptation via domain adversarial training for speaker recognition. 2018. 34, 35

[85] Sicheng Wang, Wei Li, Sabato Marco Siniscalchi, and Chin-Hui Lee. A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers. In *ICASSP*, pages 6219–6223. IEEE, 2020. 34

[86] 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. *ICASSP*, pages 2080–2091, 2019.

[87] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey. Student-teacher network learning with enhanced features. In *ICASSP*, pages 5275–5279, 2017. 34

[88] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 35

[89] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang. An unsupervised deep domain adaptation approach for robust speech recognition. *Neurocomputing*, 257:79–87, 2017. 35

[90] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. Domain adversarial training for accented speech recognition. *arXiv preprint arXiv:1806.02786*, 2018. 35

[91] Sadaoki Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, 1986. 35, 37

[92] Xiong Xiao, Shengkui Zhao, Duc Hoang Ha Nguyen, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li. Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation. *EURASIP Journal on Advances in Signal Processing*, 2016(1):4, 2016. 35, 37

[93] Chenglin Xu, Wei Rao, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Single channel speech separation with constrained utterance level permutation invariant training using grid lstm. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10. IEEE, 2018. 35, 41

[94] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. A shifted delta coefficient objective for monaural speech separation using multi-task learning. In *Proceedings of Interspeech*, pages 3479–3483, 2018. 35

[95] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46: 535–557, 2017. 39

[96] 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 *Interspeech*, pages 352–356, 2016. 39, 40, 48, 59

[97] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019. 41

[98] ITUT Rec. P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union, CH–Geneva*, 2005. 41, 48, 60

[99] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2008. 41, 48, 60

[100] Schuyler R Quackenbush, Thomas Pinkney Barnwell, and Mark A Clements. *Objective measures of speech quality*. Prentice Hall, 1988. 42

[101] Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 629–632. IEEE, 1996. 43

[102] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM TASLP*, 28:1370–1384, 2020. 45, 58

[103] Jae Lim and Alan Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(3):197–210, 1978. 50

[104] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. 2017. 50

[105] Meet H Soni, Neil Shah, and Hemant A Patil. Time-frequency masking-based speech enhancement using generative adversarial network. In *ICASSP*, pages 5039–5043. IEEE, 2018. 50

[106] Nana Hou, Chenglin Xu, Joey Tianyi Zhou, Eng Siong Chng, and Haizhou Li. Multi-task learning for end-to-end noise-robust bandwidth extension. In *INTERSPEECH*, pages 4069–4073, 2020. 53

[107] Chuping Liu, Qian-Jie Fu, and Shrikanth S Narayanan. Effect of bandwidth extension to telephone speech recognition in cochlear implant users. *The Journal of the Acoustical Society of America*, 125(2):EL77–EL83, 2009. 53

[108] Kehuang Li, Zhen Huang, Yong Xu, and Chin-Hui Lee. Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. 53

[109] Phani Sankar Nidadavolu, Cheng-I Lai, Jesús Villalba, and Najim Dehak. Investigation on bandwidth extension for speaker recognition. In *Interspeech*, pages 1111–1115, 2018.

[110] David Haws and Xiaodong Cui. Cyclegan bandwidth extension acoustic modeling for automatic speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6780–6784. IEEE, 2019. 53

[111] Ryota Kaminishi, Haruna Miyamoto, Sayaka Shiota, and Hitoshi Kiya. Investigation on blind bandwidth extension with a non-linear function and its evaluation of x-vector-based speaker verification. In *Proc. INTERSPEECH*, pages 4055–4059, 2019. 53

[112] Hitoshi Yamamoto, Kong Aik Lee, Koji Okabe, and Takafumi Koshinaka. Speaker augmentation and bandwidth extension for deep speaker embedding. *Proc. Interspeech 2019*, pages 406–410, 2019. 53

[113] Kehuang Li and Chin-Hui Lee. A deep neural network approach to speech bandwidth expansion. In *ICASSP*, pages 4395–4399. IEEE, 2015. 54, 60

[114] Johannes Abel and Tim Fingscheidt. Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1): 71–83, 2017. 54

[115] Yu Gu and Zhen-Hua Ling. Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension. In *INTERSPEECH*, pages 1123–1127, 2017. 54

[116] Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon. Audio super resolution using neural networks. *ICLR*, 2017. 54, 60

[117] Sung Kim and Visvesh Sathe. Bandwidth extension on raw audio via generative adversarial networks. *CoRR*, abs/1903.09027, 2019. URL `http://arxiv.org/abs/1903.09027`. 54

[118] Archit Gupta, Brendan Shillingford, Yannis Assael, and Thomas C Walters. Speech bandwidth extension with wavenet. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 205–208. IEEE, 2019. 54

[119] Pedro J Moreno, Bhiksha Raj, and Richard M Stern. A vector taylor series approach for environment-independent speech recognition. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 733–736. IEEE, 1996. 54

[120] Michael L Seltzer, Alex Acero, and Jasha Droppo. Robust bandwidth extension of noise-corrupted narrowband speech. In *EUSIPCO*, 2005. 54

[121] Hyunson Seo, Hong-Goo Kang, and Frank Soong. A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise. In *ICASSP*, pages 6087–6091. IEEE, 2014. 54

[122] Bin Liu, Jianhua Tao, Zhengqi Wen, Ya Li, and Danish Bukhari. A novel method of artificial bandwidth extension using deep architecture. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. 54

[123] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias icml. *Google Scholar Google Scholar Digital Library Digital Library*, 1993. 55

[124] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[125] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017. 55

[126] Yi Luo, Cong Han, Nima Mesgarani, Enea Ceolini, and Shih-Chii Liu. Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing. In *ASRU*, pages 260–267. IEEE, 2019. 56

[127] Ashutosh Pandey and DeLiang Wang. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6875–6879. IEEE, 2019. 57, 79, 85

[128] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 57

[129] 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*, pages 1026–1034, 2015. 57

[130] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. Time-domain speaker extraction network. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 327–334. IEEE, 2019. 58

[131] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019. 58, 75

[132] Xiang Hao, Chenglin Xu, Nana Hou, Lei Xie, Eng Siong Chng, and Haizhou Li. Time-domain neural network approach for speech bandwidth extension. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 866–870. IEEE, 2020. 59

[133] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 59, 78

[134] Lawrence R. Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition. *Englewood Cliffs, NJ, USA: Prentice-Hall*, 1993. 60

[135] Hou Nana, Chng Eng Siong, and Li Haizhou. Two step training for speech enhancement with multi-filterbank design. *IEEE transactions on audio, speech, and language processing*, 2021. 65

[136] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals. Learning the speech front-end with raw waveform cldnns. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. 70

[137] 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. 70

[138] Mirco Ravanelli and Yoshua Bengio. Interpretable convolutional filters with sincnet. *arXiv preprint arXiv:1811.09725*, 2018. 70

[139] Donald Eric Broadbent. *Perception and communication*. Elsevier, 2013. 71

[140] Yipeng Li and DeLiang Wang. On the optimality of ideal binary time–frequency masks. *Speech Communication*, 51(3):230–239, 2009. 71

[141] Arun Narayanan and DeLiang Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7092–7096. IEEE, 2013. 71

[142] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712. IEEE, 2015. 71, 72

[143] Emine Merve Kaya and Mounya Elhilali. Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714): 20160101, 2017. 72

[144] Emad M Grais, Gerard Roma, Andrew JR Simpson, and Mark D Plumbley. Two-stage single-channel audio source separation using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9): 1773–1783, 2017. 74

[145] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 74

[146] Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2018. 75

[147] Chandan KA Reddy, Ebrahim Beyrami, Harishchandra Dubey, Vishak Gopal, Roger Cheng, Ross Cutler, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework. *arXiv preprint arXiv:2001.08662*, 2020. 76

[148] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, pages 146–152, 2016. 76

[149] Jodi Kearns. Librivox: Free public domain audiobooks. *Reference Reviews*, 2014. 76

[150] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 76

[151] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017. 76

[152] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël AP Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, et al. A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016(1):1–19, 2016. 76

[153] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. Multichannel audio database in various acoustic environments. In *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 313–317. IEEE, 2014. 76

[154] Yangyang Xia, Sebastian Braun, Chandan KA Reddy, Harishchandra Dubey, Ross Cutler, and Ivan Tashev. Weighted speech distortion losses for neural-network-based real-time speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 871–875. IEEE, 2020. 77, 79, 85

[155] Xiaofei Li and Radu Horaud. Online monaural speech enhancement using delayed subband lstm. *arXiv preprint arXiv:2005.05037*, 2020. 77, 79, 85

[156] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001. 80