DNN-Assisted Speech Enhancement Approaches Incorporating Phase Information

Hongjiang Yu

A Thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Electrical and Computer Engineering) at
Concordia University
Montréal, Québec, Canada

April 2021

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Hongjiang Yu**

Entitled: **DNN-Assisted Speech Enhancement Approaches Incorporating Phase Information**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Chair
Dr. Anjan Bhowmick

______________________________ External Examiner
Dr. Xiao-Ping Zhang

______________________________ External to Program
Dr. Adam Krzyzak

______________________________ Examiner
Dr. Omair Ahmad

______________________________ Examiner
Dr. Hassan Rivaz

______________________________ Thesis Supervisor
Dr. Wei-Ping Zhu

Approved by ______________________________
Dr. Wei-Ping Zhu, Graduate Program Director

April 27, 2021 ______________________________
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

DNN-Assisted Speech Enhancement Approaches
Incorporating Phase Information

Hongjiang Yu, Ph.D.

Concordia University, 2021

Speech enhancement is a widely adopted technique that removes the interferences in a corrupted speech to improve the speech quality and intelligibility. Speech enhancement methods can be implemented in either time domain or time-frequency (T-F) domain. Among various proposed methods, the time-frequency domain methods, which synthesize the enhanced speech with the estimated magnitude spectrogram and the noisy phase spectrogram, gain the most popularity in the past few decades. However, this kind of techniques tend to ignore the importance of phase processing. To overcome this problem, the thesis aims to jointly enhance the magnitude and phase spectra by means of the most recent deep neural networks (DNNs). More specifically, three major contributions are presented in this thesis.

First, we present new schemes based on the basic Kalman filter (KF) to remove the background noise in the noisy speech in time domain, where the KF acts as joint estimator for both the magnitude and phase spectra of speech. A DNN-augmented basic KF is first proposed, where DNN is applied for estimating key parameters in the KF, namely the linear prediction coefficients (LPCs). By training the DNN with a large database and making use of the powerful learning ability of DNN, the proposed algorithm is able to estimate LPCs from noisy speech more accurately and robustly, leading to an improved performance as compared to traditional KF based approaches in speech enhancement. We further present a high-frequency (HF) component restoration algorithm to extenuate the degradation in the HF regions of the Kalman-filtered speech, in which the

DNN-based bandwidth extension is applied to estimate the magnitude of HF component from the low-frequency (LF) counterpart. By incorporating the restoration algorithm, the enhanced speech suffers less distortion in the HF component. Moreover, we propose a hybrid speech enhancement system that exploits DNN for speech reconstruction and Kalman filtering for further denoising. Two separate networks are adopted in the estimation of magnitude spectrogram and LPCs of the clean speech, respectively. The estimated clean magnitude spectrogram is combined with the phase of the noisy speech to reconstruct the estimated clean speech. A KF with the estimated parameters is then utilized to remove the residual noise in the reconstructed speech. The proposed hybrid system takes advantages of both the DNN-based reconstruction and traditional Kalman filtering, and can work reliably in either matched or unmatched acoustic environments.

Next, we incorporate the DNN-based parameter estimation scheme in two advanced KFs: subband KF and colored-noise KF. The DNN-augmented subband KF method decomposes the noisy speech into several subbands, and performs Kalman filtering to each subband speech, where the parameters of the KF are estimated by the trained DNN. The final enhanced speech is then obtained by synthesizing the enhanced subband speeches. In the DNN-augmented colored-noise KF system, both clean speech and noise are modelled as autoregressive (AR) processes, whose parameters comprise the LPCs and the driving noise variances. The LPCs are obtained through training a multi-objective DNN, while the driving noise variances are obtained by solving an optimization problem aiming to minimize the difference between the modelled and observed AR spectra of the noisy speech. The colored-noise Kalman filter with DNN-estimated parameters is then applied to the noisy speech for denoising. A post-subtraction technique is adopted to further remove the residual noise in the Kalman-filtered speech. Extensive computer simulations show that the two proposed advanced KF systems achieve significant performance gains when compared to conventional Kalman filter based algorithms as well as recent DNN-based methods under both seen and unseen noise conditions.

Finally, we focus on the T-F domain speech enhancement with masking technique, which aims to retain the speech dominant components and suppress the noise dominant parts of the noisy

speech. We first derive a new type of mask, namely constrained ratio mask (CRM), to better control the trade-off between speech distortion and residual noise in the enhanced speech. The CRM is estimated with a trained DNN based on the input noisy feature set and is applied to the noisy magnitude spectrogram for denoising. We further extend the CRM to the complex spectrogram estimation, where the enhanced magnitude spectrogram is obtained with the CRM, while the estimated phase spectrogram is reconstructed with the noisy phase spectrogram and the phase derivatives. Performance evaluation reveals our proposed CRM outperforms several traditional masks in terms of objective metrics. Moreover, the enhanced speech resulting from the CRM based complex spectrogram estimation has a better speech quality than that obtained without using phase reconstruction.

Acknowledgments

First and foremost, I would like to express my sincerest gratitude and appreciation to my supervisor, Prof. Wei-Ping Zhu, for leading me to the area of speech enhancement, teaching me the basic knowledge as well as the state-of-art techniques, supporting me when I was obstructed in the research and encouraging me when I encountered difficulties.

Besides, I want to give my special thanks to Prof. Benoit Champagne, McGill University, Canada. He gives lots of supports and advice about my research project and publication, which develops my theoretical knowledge in speech processing and improves my writing skills in technical papers.

My heartiest thanks also go to Concordia for providing me the opportunity to study in such a great university, to Microchip in Ottawa for sponsoring our NSERC CRD research project, and to the financial support from China Scholarship Council.

I would like also extend my deep gratefulness to all the members in our project for their collaborations and feedbacks on my presentations during the research meetings. I am also grateful to my group members, Mr. Xinrui Pu, Mr. Mojtaba Hasannezhad, Mr. Zhiheng Ouyang and all the laboratory members and friends for their help and comforts throughout my life during the past five years.

I also owe many thanks to the professors in my examining committee for their guidances and comments to my comprehensive exam, proposal, seminar and thesis.

Finally, I would like to express my deepest love to my parents. Their selfless love, encouragement and support are always the source of my strength to overcome all the frustrations and obstacles in my life.

Contents

List of Figures

List of Tables

List of Abbreviations

AMS Amplitude Modulation Spectrum

AR Autoregressive

cIRM Complex Ideal Ratio Mask

CNN Convolutional Neural Network

CRM Constrained Ratio Mask

DCN Dense Convolutional Network

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DNN Deep Neural Network

DWT Discrete Wavelet Transform

FNN Fully-connected Feed-forward Deep Neural Network

GAN Generative Adversarial Network

GD Group Delay

GFCC Gammatone Frequency Cepstral Coefficients

HF High-Frequency

IAM Ideal Amplitude Mask

IBM Ideal Binary Mask

iDWT Inverse Discrete Wavelet Transform

IF Instantaneous Frequency

IFD Instantaneous Frequency Deviation

IRM Ideal Ratio Mask

iSTFT Inverse Short-Time Fourier Transform

KF Kalman Filter

LF Low-Frequency

LPCs Linear Prediction Coefficients

LPS Log-Power Spectra

LSFs Line Spectrum Frequencies

LSTM Long Short-Term Memory

MFCC Mel-Frequency Cepstral Coefficients

MMSE Minimum Mean Square Error

MOS Mean Opinion Score

MRCG Multi-Resolution Cochleagram

MSE Mean Square Error

ORM Optimal Ratio Mask

PESQ Perceptual Evaluation of Speech Quality

PLP Perceptual Linear Prediction

PNCC Power Normalized Cepstral Coefficients

PSM Phase Sensitive Mask

RASTA Relative Spectral Transform

ReLU Rectified Linear Unit

RGAN Relativistic Generative Adversarial Network

RI Real and Imaginary

RNN Recurrent Neural Network

SDR Source-to-Distortion Ratio

SE Speech Enhancement

SegSNR Segmental Signal-to-Noise Ratio

SNR Signal-to-Noise Ratio

STFT Short-Time Fourier Transform

STOI Short-Time Objective Intelligibility

STSA Short-Time Spectral Amplitude

T-F Time-Frequency

VAD Voice Activity Detection

ZCPA Zero-Crossings with Peak-Amplitudes

Chapter 1

Introduction

In this chapter, we first present a brief introduction of speech enhancement (SE), and its practical applications. Next, a general literature review on the existing methods in both time domain and time-frequency (T-F) domain is followed. The motivation and objectives of this research are discussed in the subsequent section. At the end, a chapter-by-chapter organization of this thesis and the major contributions are described.

1.1 Speech Enhancement and Its Applications

Speech signal is the most common and convenient carrier in communication. However, in real-world environments, clean speech is often corrupted by a wide range of interferences, which causes the degradation of speech quality and further harms the user experiences in speech communication applications. SE, which aims to remove the corrupted interferences and improve speech quality and intelligibility, has been intensively studied over the past several decades, and will likely continue to be an active topic in speech processing, especially with the development of the artificial intelligence (AI).

SE plays an important role in a wide variety of applications including mobile phones, voice over internet protocol, teleconferencing systems, speech recognition, and hearing aids [1]. For instance, voice communication systems over cellphones suffer from distortion when transmitting in noisy environment. SE algorithms are therefore used to improve the quality of speech at the receiving end. Moreover, the ambient noises in a teleconferencing system will be captured by microphones and then transmitted to all the receivers. Enhancing the noisy speech before transmission will be desired to obtain better performance. Finally, SE is also vital to hearing aid devices to help hearing-impaired listeners gain better communication experience in noisy conditions by pre-processing the noisy signal before amplification. Thus, with the fast development of the aforementioned speech and audio systems, there is a growing need for further development of SE algorithms in the future.

SE can be categorized from different aspects. According to the number of the noisy speech channel, SE can be divided into two categories: single-channel (monaural) and multi-channel. As well-known, monaural SE is a more challenging problem since we can not access additional information from other channels. SE can also be classified into different subtasks, i.e., denoising, separation, and dereverberation. Denoising is to remove the background noise from a noisy speech signal, while separation aims to extract several different source speech signals from a mixture signal. Unlike speech denoising or separation, where the desired speech is corrupted by other interferences, dereverberation is to deal with the echo or late reverberation from the speech signal itself. The SE problem considered in this paper is single-channel speech denoising, that is, to recover the desired speech signal from the noisy observation, which can be performed in either time domain or T-F domain.

1.2 Speech Enhancement in Time Domain

In early works, on time domain SE, an appropriate filter is designed and applied to the noisy observation to obtain the desired signal [2]. Wiener filter is one of the most famous filters, which is derived by finding the optimal minimum mean square error (MMSE) estimate of the clean speech [3, 4]. The performance of Wiener filtering is limited due to the impractical assumptions. For instance, the enhanced speech suffers from musical noise when dealing with the non-stationary noise, as both the speech signal and noise are assumed to be stationary in the derivation. Kalman filter (KF), which can handle non-stationary signals, has therefore attracted the interests of researchers [5] in SE area. In this context, the KF can be viewed as a time domain, sequential linear MMSE estimator of the noise corrupted speech, in which the clean speech is characterized by a dynamical or state-space model, such as the autoregressive (AR) model. As such, the enhancement performance is largely dependent on the estimation accuracy of the AR parameters, which include the LPCs and the variances of the driving and observation noises.

Recently, there has been a great deal of interest in data-driven supervised methods for SE. More specifically, a DNN is used as an encoder-decoder that takes the noisy speech waveform as input, and outputs the waveform of the clean speech. However, the authors in [6] pointed out that the fully-connected feed-forward neural network (FNN) is not suitable for time domain SE, because the sample point does not contain much information. Instead, a fully convolutional neural network (CNN) is investigated to estimate the clean speech waveform, as the convolution operation can efficiently find useful locally acoustic information. Based on this finding, various CNN-based frameworks have been proposed for time domain SE in the past few years [6–11]. The multi-resolution convolutional auto-encoder (MRCAE) [8] consists of two convolution and transposed convolution layers, and uses different convolutional filter sizes to detect audio frequencies with different resolutions. But MRCAE can only extract little context information as the network does not perform any resampling and takes only the speech with one time-resolution as input. Another popular framework is Wave-U-Net with several downsampling blocks in the encoder, and upsampling blocks in the decoder [12], which can calculate the longer-term dependencies based on feature maps with more lower-resolution features. More recently, the authors in [11] have proposed a dense convolutional network (DCN) with self-attention for SE in the time domain, based on the finding that a SE system with self-attention can better reconstruct the enhanced speech based on the local signal-to-noise ratio (SNR) of different regions.

Moreover, the deep generative models have also been applied to time domain SE. One of the pioneering works is the SE generative adversarial network (GAN) proposed in [13], in which the generator is an auto-encoder based fully-convolutional network that is trained with the help of a discriminator. More specifically, the generator takes the raw noisy speech as input and generates the estimated clean speech, which is similar to the ideal clean speech as much as possible, while the discriminator attempts to distinguish whether the input speech is the estimated one from the generator or ideal one from the dataset. Experimental results in [13] show that the SEGAN outperforms the Wiener filtering and statistical model based methods. However, the performance of SEGAN system is still not good enough in the situation where the training data is not sufficient. In this case, the instability of training and the problem of gradient disappearing result in an inadequate training process. As such, several improved GAN systems have been proposed in recent works [14–16]. In [14], the Wasserstein conditional GAN with gradient penalty is proposed to improve the performance of the model when large traing datasets are not satisfied. In [15], the authors introduced the relativistic GAN (RGAN) that uses a relativistic loss function at the discriminator, and investigate whether RGAN can yield a better generator network for SE. The simulation results show that the RGAN has a more stable training process and better performance than the SEGAN. In [16], the authors combined the Wave-U-Net with GAN and proposed the UNetGAN. By employing adversarial learning to improve the U-Net in the time domain, the UNetGAM significantly improves the speech quality and achieves state-of-the-art performance at extremely low SNR conditions.

Another widely-used generative model is WaveNet proposed by Google [17], which is first applied in speech synthesis to generate raw audio waveforms. WaveNet is then utilized for SE in [18], in which the model takes the noisy speech waveform as input and outputs the estimated clean speech waveform. One thing to mention is that although the WaveNet is a generative model, it is discriminative when used in SE as its output is not explicitly modeling a probability distribution. The experimental results show that the denoising performance of WaveNet is better than Wiener filtering.

In conclusion, the time domain SE has attracted much attention in the early days, since it is an end-to-end process that avoids the computation complexity of Fourier transform. However, the time domain waveform does not make full use of the acoustic features in the T-F spectrogram. Therefore, more algorithms are developed in enhancing the short-time Fourier transform (STFT) spectrogram within the past decades. Recently, the powerful learning capability of the deep model offers an opportunity to extract information from the speech waveform. As such, the time domain SE gains the researchers' interests again and achieves breakthroughs in recent years.

1.3 Speech Enhancement in Time-Frequency Domain

SE in T-F domain is a popular branch of methods as the STFT spectrogram displays speech information in a different view compared to the raw waveform. Since a STFT spectrogram consists of both magnitude and phase, the following section will introduce the denoising of the magnitude and the phase spectra separately.

1.3.1 Magnitude Processing

Most previous denoising algorithms are focused only on processing magnitude spectrograms, among which spectral subtraction [19] is the most intuitive one. The main idea of this method is to estimate the noise spectrum in the speech-absent segments and subtract it from the noisy speech. Although spectral subtraction is easy to employ, estimating an accurate noise spectrum is a difficult task. Either over-estimated or under-estimated could bring extra distortions to speech, such as the notorious musical noise, which might be more annoying than the background noise. Although spectral subtraction is intrusive and easy to employ, the difficulty in estimating accurate noise spectrum hinders the enhancement performance. More flexible spectral subtraction algorithms with better performance are proposed in [20, 21], where two techniques, i.e., the use of oversubtraction factor and spectral flooring parameter, were introduced along with the standard spectral subtraction. These techniques are used to adjust the estimated noise spectrum, and thereby control the ratio of the remaining residual noise and perceived musical noise in the enhanced speech. In [22, 23], a multiband spectral subtraction has been proposed based on the fact that the noise affects the speech at different levels depending on frequency bands. In the multiband approach, the speech spectrum is divided into several non-overlapping frequency bands, and then spectral subtraction is performed independently in each band.

Moreover, the time domain statistical filtering methods can be extended to T-F domain. For instance, Wiener filtering[24, 25], which aims to find the optimal MMSE estimate of the discrete Fourier transform (DFT) coefficients of the clean speech, obtains relatively better results compared with spectral subtraction. However, the filtered speech still suffers distortion as Wiener filtering introduces residual noise instead of musical noise in the enhanced speech. Another example is the modulation domain Kalman filtering in [26], where the noisy magnitude spectrum is viewed as a series of modulating signals that span across time. The KF is then applied to each modulating signal to estimate the clean speech magnitude spectrum. Subjective listening tests demonstrate that the modulation domain KF outperforms the time domain counterpart in terms of speech quality.

Unlike Wiener filter which aims to find optimal complex spectral estimator, the short-time spectral amplitude (STSA) estimators focus on obtaining the optimal spectral amplitude estimator [27, 28]. A Bayesian framework is employed to derive the STSA estimators, based on the assumptions about the probability distributions of speech and noise DFT coefficients. Early STSA estimators used a Gaussian assumption in derivation, then a large range of estimators have been proposed to improve the enhancement performance either with more accurate statistical assumptions that are more related with the true probability distribution of speech and that of noise [29–31]. Although the STSA estimators can substantially reduce the residual noise, similarly to the Wiener filters, a priori SNR has to be estimated in order to use the STSA estimators in practice.

Instead of using STSA estimators to predict the clean speech magnitude spectrum directly, other kinds of techniques suppress the background noises in the noisy speech magnitude spectrum. Among them, T-F masking is a famous example, which was first proposed in computational auditory scene analysis to separate speech from noisy mixtures by Wang [32]. Inspired by the masking effects of human auditory system, an estimated mask is applied to each T-F units of the noisy speech, in order to conserve the speech-dominant region and suppress the noise-dominant region. Ideal binary mask (IBM) is the first one in masking techniques [32, 33], which is defined as a binary matrix with 1 denoting that the speech energy in the corresponding T-F unit exceeds the noise energy by a predefined threshold and 0 denoting the opposite. The enhanced T-F unit is obtained by applying the estimated IBM to the noisy speech spectrogram. However, IBM is not accurate enough as it is a hard-decision mask. Ideal ratio mask (IRM) [34], as a soft-decision mask, is then proposed to better suppress the background noise in the T-F unit, providing a better enhancement performance than IBM.

The recent deep learning techniques are also widely utilized for T-F domain SE. In [35], Xu et. al have established a regression model to directly learn the mapping between the log-power spectra (LPS) of noisy speech and that of clean speech based on DNN. By training the DNN with a large set that encompasses many possible combinations of speech and noise types. The estimated magnitude spectrum is transformed from the estimated LPS, which is given by the well-trained DNN. The enhanced speech is then reconstructed from the estimated magnitude and the noisy phase, leading to significant improvements in terms of both objective and subjective measures over the conventional MMSE-STSA estimator.

DNN has also been used as a primary tool to predict key parameters in traditional SE methods. For example, Wang et. al in [36] employ DNN to estimate the IRM for masking based algorithms. The enhanced speech is then obtained by applying the estimated mask to the noisy speech. In [37], a DNN is trained to estimate the LPCs of both the clean speech AR model and the noise AR model given the noisy observation. Then, a Wiener filter is constructed with the estimated LPCs, which is then applied to the noisy speech spectrogram for noise reduction. In [38], a DNN-based denoising method is proposed in order to use a harmonic noise model for the estimation of the clean speech amplitude, where the DNN is employed to map noisy speech features to clean amplitude parameters of the harmonic noise model. More recently, a long short-term memory (LSTM) network is utilised in [39] to accurately estimate the *a priori* SNR for traditional MMSE-STSA estimators. Because of the powerful learning ability of the DNN, researchers can obtain the more accurate key parameters estimates, thus the aforementioned DNN-assisted traditional SE methods achieve better performances in terms of both the objective quality and intelligibility scores than their counterparts without deep learning. Another breakthrough includes the deep learning based generative modeling for SE, wherein the GAN has been successfully employed to generate clean speech magnitude spectrogram [40, 41] with the acoustic features of the noisy speech as input.

In the past few decades, researchers have made huge efforts in enhancing the magnitude spectrogram and achieved significant improvement. However, a common problem of these methods is that the noisy phase is directly used in the reconstruction of the enhanced speech in most T-F methods.

1.3.2 Phase Processing

Mainstream approaches have tended to ignore the phase processing for two reasons. Firstly, researchers hold the opinions that our ears are insensitive to small phase distortions [42] in early works, which indicates that the enhancement performance would not decrease too much when using noisy phase for reconstruction. Secondly, unlike the magnitude spectrum, which has a clear harmonic structure to be estimated, the processing of the phase remains challenging due to its unstructured characteristic and phase wrapping [43]. As the performance of magnitude-only methods is limited without considering phase, and as the computational power of speech communication devices, reinvestigation of phase processing in SE is back to the researchers' sight. Some researchers have pointed out the importance of estimating clean phase in recent works, especially at low SNRs [44, 45]. For instance, Paliwal et al. [45] reported that when combining an MMSE estimate of the clean speech magnitude with the oracle clean speech phase in a perfectly reconstructing STFT framework, an improvement of 0.2 points of the mean opinion score (MOS) is observed by the perceptual evaluation of speech quality (PESQ) measure for white Gaussian noise. The research confirms the importance of developing and improving phase processing algorithms. Therefore, phase enhancement has recently been the focus of multiple research groups.

Among the first proposals for phase estimation are the iterative approaches, which aim at estimating a time domain signal whose STFT magnitude is as close as possible to a target one. The most well-known and fundamental approach in this category is that of Griffin and Lim [46], which applies STFT synthesis and analysis iteratively while retaining the information about the updated phases and replacing the updated magnitudes with the target magnitudes. This exploits correlations between neighboring STFT frames to obtain an estimate of the spectral phases and the time domain signal. One of the problems with this method is its high computational complexity. A solution is proposed by Le Roux et al. [47] based on the standard operation of classical iterative approaches, i.e., computing the STFT of the signal obtained by inverse STFT (iSTFT) from a given spectrogram, can indeed be considered as a linear operator in the T-F domain. Le Roux et al. noticed that the result of that operation at each T-F bin can be well approximated by a local weighted sum with complex coefficients on a small neighborhood of that bin in the original spectrogram. Although Le Roux's method is conceptually close to the iterative STFT algorithm introduced by Griffin and Lim, the computational cost is reduced by employing on local phase coherence conditions and enabling at each iteration the update of each T-F bin's phase independently.

In contrast to the iterative approaches, sinusoidal model-based phase estimation does not require estimates of the clean speech spectral magnitudes. Instead, the clean spectral phase is estimated using only an estimate of the fundamental frequency, which can be obtained from the degraded signal. In [48], Krawczyk et al. proposed a method to recover the clean spectral phase of voiced speech along time and frequency with the employment of harmonic model, where the spectral phase is estimated between the harmonic components. They reported that this phase reconstruction between the harmonics achieves better noise reduction during voiced speech when the phase estimates are employed for SE. Informal listening confirms the noise reduction during voiced speech at the expense of a slightly audible residual signal, which can be effectively alleviated by combination with amplitude enhancement. However, since the usage of the sinusoidal model is reasonable only for voiced sounds, these approaches do not provide valid spectral phase estimates for unvoiced sounds, like fricatives or plosives.

Besides the value for signal reconstruction, phase estimation can also be utilized as additional information for phase-aware magnitude estimation. For example, Gerkmann and Krawczyk [49] have derived an MMSE estimator of the spectral magnitude given the clean speech phase. Experiments demonstrate that an improved magnitude estimator derived with the information of the speech spectral phase can reduce noise outliers that are not tracked by the noise power spectral density estimator. In [50], the authors suggested incorporating phase estimation during signal reconstruction to improve the quality of the T-F masked separation outcome when applied on mixture signals. The proposed method replaces the mixed spectral phase with an estimated clean spectral phase, which is used for the reconstruction of the separated signals. The estimated spectral phase is calculated by temporal smoothing of the unwrapped phase, which is provided by harmonic phase decomposition of the mixture phase given the fundamental frequency of the target signal. The experiments demonstrate that replacing the mixture phase with the estimated clean spectral phase consistently improves perceptual speech quality, predicted speech intelligibility, and source separation performance across all SNRs and different noise scenarios.

Instead of estimating phase and magnitude separately, a better choice is to estimate them jointly to explore the relationship between magnitude and phase spectra. The first step in this direction is to incorporate the phase information as a constraint in deriving the statistical based filters or STSA estimators, such as the consistent Wiener filter proposed by Le Roux and Vincent [51]. As a classical Wiener filter only changes the magnitudes in the STFT domain, the modified spectrum is inconsistent. In contrast to this, the consistent Wiener filter considers the relationship between STFT coefficients across time and frequency, which modifies both the magnitude and the phase of the noisy observation to obtain the enhanced speech. With the joint estimation, the consistent Wiener filter was shown to lead to an improved enhancement performance compared to the classical Wiener filter. In [52], the authors incorporated the phase information to masking technique and proposed a phase sensitive mask (PSM). Experiments show that the speech denoised by the estimated PSM achieves better performance in the subsequent speech recognition task than the one by IRM.

Another way to jointly enhance magnitude and phase is to process the real and imaginary (RI) spectrograms of the STFT domain signal, and the enhanced speech is obtained by synthesizing with the enhanced RI parts. In [53, 54], the authors used a DNN to learn a spectral mapping from noisy speech to a complex IRM (cIRM), which is then applied to the noisy speech in the RI domains for noise reduction. Results show that the estimated cIRM substantially outperforms directly estimating speech in the time domain, traditional IRM estimation in the magnitude domain. Furthermore, cIRM estimation is shown to outperform methods that separately enhance the magnitude and phase of noisy speech. In [55], a novel CNN model is proposed for complex spectrogram estimation, which employs CNN for extracting the latent features of the RI spectrograms and uses an FNN as a regression model to estimate clean RI spectrograms from noisy ones. The estimated RI spectrograms are directly used to reconstruct the enhanced speech waveforms. Experimental results confirm the effectiveness of the proposed CNN with RI spectrograms on a SE task. More recently, the authors further improved the CNN-RI framework by adopting a fully dilated CNN for complex spectrogram estimation [56].

The aforementioned DNN-based methods do not directly tackle the difficulty of processing a phase spectrogram, as the phase spectrogram is randomly distributed and highly unstructured. Even for the DNN, a direct mapping from the noisy phase to the clean phase would not be easy. As a result, the alternative representations of the phase have been proposed to reveal the structure of the phase, by considering the relationships between neighboring T-F units. In [57], the derivative of the STFT phase, namely, instantaneous frequency deviation (IFD), is used as a training target of the DNN. The estimated IFD is then converted back to the phase for speech reconstruction. However, the transformation between IFD and phase increases the computational complexity of the approach. In [58], a two-stream network with two-way information exchange named PHASEN is proposed, where amplitude stream and phase stream are dedicated to amplitude and phase prediction. The authors concluded that the phase prediction would be improved a lot if two streams communicate with each other.

Current researches reveal that considering the phase information of speech signals allows one to achieve better enhancement performance. However, such studies are still in their preliminary stages. It is well known that incorporating perceptual rules into the magnitude processing has received considerable attention in the relevant literature [29], but only a few improvements and modifications of perceptually and simultaneously enhancing magnitude and phase have been presented so far.

1.4 Motivations and Objectives of the Research

1.4.1 Motivations

From the above literature overview, design of new SE methods with neural networks is desperately needed for the rapidly growing market of speech and audio processing applications. Even though SE has been extensively studied over the past three decades, there are still several uncertain and unresolved issues in this area. In this section, we summarize the motivation behind this research from the following perspectives.

Time domain vs. T-F domain: As presented in Section 1.2 and Section 1.3, SE can be performed in either time domain or T-F domain. However, the vast majority of studies are conducted in T-F domain as the acoustic features extracted from STFT spectrogram of a speech signal leads to a good performance in SE. Early methods in T-F domain only processed the magnitude spectrogram, while complex spectrogram enhancement has attracted the researcher's interest due to the incorporation of phase information. Even though complex spectrogram enhancement and time domain enhancement have similar objectives, the latter catches the researchers' attention for several reasons [11]. First, time domain enhancement avoids the computational complexity associated with the STFT and iSTFT. Second, DNN has the potential to extract better and more suitable features for the particular task of SE when trained with raw speech waveforms. Finally, short-time processing in T-F domain requires frame size to be greater than a certain threshold to have sufficient spectral resolution, whereas in time domain processing frame size is more flexible, which can be set to an arbitrary value. As time domain enhancement and T-F domain enhancement have their own advantages, the first motivation of this research is to investigate the enhancement methods in each domain and further to make use of both advantages.

Unsupervised vs. supervised: In recent years, deep learning, and especially DNN, has been successfully applied in SE and quickly becomes one of the most popular techniques. Compared with the unsupervised statistical-model based methods, the use of DNN offers several advantages. First, the non-linear structures of DNN confer them with powerful learning capability, suitable to model the complex mapping relationship between the noisy and clean speech. Furthermore, deep learning based methods usually do not require the estimation of the noise power spectrum, nor do they rely on particular assumptions about the statistical properties of the speech and noise, which allow them to handle non-stationary noises in real-world scenarios under unexpected acoustic conditions. However, deep learning based algorithms require large databases for training in order to improve their generalization capability. To achieve better performance in unseen noise conditions, it is common to train a DNN with a large speech database comprising different speakers and noise types [35].

Although the conventional unsupervised statistical model based methods fail to achieve satisfactory results in real-world environments, the fact that they can reduce different kinds and levels of noises to some extent, is attractive to researchers. In other words, the statistical model based methods do not employ a training stage, and thus treat all noises as unseen noise so that their denoising capability, albeit limited, remains available in all situations. Based on such considerations, the development of hybrid approaches, which take advantage of both unsupervised methods and deep learning methods, is regarded as another major motivation of this research.

Residual noise vs. speech distortion: In ideal case, to obtain high-quality enhanced speech, one should remove the background noise as much as possible while do not bring distortion to the speech. However, such distortion is inevitable in practice as the speech information receives damage along with the noise reduction. Therefore, the trade off between residual noise and speech distortion has always been an important topic in SE over the past few decades. Existing methods such as constrained Wiener filtering [59] and perceptual STSA estimators [60] deal with the problem by incorporating the perception rules in their derivations to exploit the masking properties of the human auditory system. In the regions where speech energy is high, these methods attenuate noise reduction to preserve the speech information, since the background noise is less likely to be audible in this case. In contrast, the denoising will be strengthened in the speech absent regions or the regions with small speech energy, so that the residual noise will be maximally removed. For recent supervised methods, the control of residual noise is realized by modifying the loss function of the DNN [61]. The perceptually weighted mean square error (MSE) [62–65] and the metric based loss function [66, 67] are two of the most popular categories. The former helps the DNN adjust the level of the noise reduction in the T-F regions according to the perceptual rules, while the latter attempts to obtain better optimization of the specific objective metric to control the speech quality and intelligibility. In order to obtain better enhancement performance, making a balance between noise attenuation and speech distortion is worth to be investigated as part of this work.

1.4.2 Objectives

Since recent works demonstrate that incorporating phase information into SE is beneficial to obtain better performance, jointly enhancing both magnitude and phase spectra is the primary goal in this research, which is investigated from two aspects: time domain Kalman filtering and T-F domain Masking. The main objectives in each aspect are summarized as follows:

- As mentioned in Section 1.2, the performance of the traditional unsupervised time domain KF is limited due to the difficulty of estimating AR parameters from noisy observation. We therefore propose a DNN-augmented basic Kalman filter, where the DNN is employed to improve the accuracy of the estimated parameters. However, the enhanced speech from the basic KF suffers the speech distortion in the high-frequency (HF) component. We then propose a restoration technique to compensate the HF component. In addition, to investigate the benefits of combining time domain enhancement and T-F domain enhancement, a hybrid system of DNN-based speech reconstruction and Kalman filtering is proposed. At last, based on the fact that the advanced KFs outperform the basic KF, we improve the DNN-augmented basic KF framework by substituting the basic KF with two advanced versions: the subband KF and the colored-noise KF, respectively. The former aims to better remove the noise according to the noise level in the subbands, while the latter considers both clean speech and noise as AR processes, and avoids the estimation additive noise variance in the basic KF.

- In the T-F domain, the well-known DNN-assisted masking technique is investigated. Although the current masks, such as IBM and IRM, are proposed to simulate the masking effects of human auditory system, most of them simply suppress the background noise according local SNR and do not take the perception principles into account. To address this deficiency, we first introduce speech and noise distortions as constraints in the derivation of the ratio mask; that is, to derive an optimal CRM that controls the trade-off between the speech distortion and residual noise. Moreover, as the proposed CRM aims to enhance the magnitude spectrogram only, we further extend the CRM to the complex spectrogram estimation to jointly enhance both magnitude and phase spectra.

1.5 Publications from the Thesis Research

Throughout the research work that has originated this thesis, the following peer-reviewed conference and journal papers have been published/submitted:

[1] **H. Yu**, W.-P. Zhu, and B. Champagne. “Speech Enhancement Using a DNN-Augmented Colored-Noise Kalman Filter.” *Speech Communication*, vol. 125, pp142-151, 2020.

[2] **H. Yu**, Z. Ouyang, W.-P. Zhu and B. Champagne. “A Hybrid Speech Enhancement System with DNN Based Speech Reconstruction and Kalman Filtering.” *Multimedia Tools and Applications*, vol. 79, pp. 32643-32993, 2020.

[3] **H. Yu**, W.-P. Zhu, and B. Champagne. “Subband Kalman Filtering with DNN Estimated Parameters for Speech Enhancement.” *Proc. of INTERSPEECH*, pp. 2697-2740, 2020.

[4] **H. Yu**, W.-P. Zhu, and Y. Yang. “Constrained Ratio Mask for Speech Enhancement Using DNN.” *Proc. of INTERSPEECH*, pp.2427-2431, 2020.

[5] **H. Yu**, W.-P. Zhu, and B. Champagne. “High-frequency Component Restoration for Kalman

Filter Based Speech Enhancement." *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, pp.1-5, 2020.

[6] **H. Yu** and W.-P. Zhu, "Deep Neural Network based Complex Spectrogram Reconstruction for Speech Bandwidth Expansion," *IEEE Int. New Circuits and Systems Conf. (NEWCAS)*, pp. 110-113, 2020.

[7] **H. Yu**, Z. Ouyang, W.-P. Zhu, B. Champagne, and Y. Ji. "A Deep Neural Network Based Kalman Filter for Time Domain Speech Enhancement." *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, pp. 1-5. 2019.

[8] Z. Ouyang, **H. Yu**, W.-P. Zhu and B. Champagne. "A Fully Convolutional Neural Network for Complex Spectrogram Processing in Speech Enhancement." *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP*), pp. 5756-5760. 2019.

[9] Z. Ouyang, **H. Yu**, W.-P. Zhu and B. champagne. "A Deep Neural Network Based Harmonic Noise Model for Speech Enhancement." *Proc. of INTERSPEECH*, pp. 3224-3228. 2018.

[10] M. Hasannezhad, **H. Yu**, W.-P. Zhu, and B. Champagne. "PACDNN: A Phase-Aware Composite Deep Neural Network for Speech Enhancement." *Speech Communication*, 2021 (Submitted).

[11] **H. Yu**, M. Hasannezhad, W.-P. Zhu, and B. Champagne. "Constrained Ratio Mask for Complex Spectrogram Estimation in Speech Enhancement." 2021. (To be submitted).

1.6 Organization

The thesis is organized as follows. In Chapter 2, a more detailed background on the introduced topics in this section is presented, based on which new speech enhancement approaches are developed and evaluated in later chapters, including Kalman filtering-based methods which constitute the main contributions in chapters 3 and 4, and Masking techniques in chapter 5. Conclusions are finally presented in Chapter 6. A detailed structure of this thesis is explained below.

In Chapter 2, a background on the topic of SE with a focus on DNN-based methods is presented. In Section 2.1, we explain the noisy speech model in both time and T-F domains. In Section 2.2, we introduce several acoustic features which have been widely applied to form the input feature set in DNN-based SE methods. In Section 2.3, an overview of popular DNN architectures for SE is briefly presented and the selection of hyper parameters for better performance is discussed. Finally, in Section 2.4, several objective metrics are presented for the evaluation of the enhanced speech.

In Chapter 3, we first present the DNN-augmented basic KF for time domain SE in Section 3.2. Next, an HF component restoration scheme is proposed in Section 3.3 as a post-processing of the KF based SE system. Finally, a hybrid system that combines DNN-based speech reconstruction with Kalman filtering technique is proposed in Section 3.4 to take advantages of both techniques. Finally, the performance evaluation of the proposed systems is given in Section 3.5.

In Chapter 4, we apply the DNN-based LPCs estimation in the previous chapter in the case of advanced KFs. More specifically, a subband KF with DNN-estimated parameters is proposed in Section 4.2, while a DNN-augmented colored-noise KF is introduced in Section 4.3. The experimental results of these two methods are shown in Section 4.4.

Chapter 5 is devoted to the T-F domain SE with masking techniques. In this respect, we derive a CRM in Section 5.2 to enhance the magnitude spectrogram of the noisy speech. The CRM is then extended to the complex spectrogram estimation in Section 5.3. At last, we compare the proposed new systems with several traditional masking methods in Section 5.4.

In Chapter 6, we draw some concluding remarks highlighting the main contributions of this thesis, and based on this we suggest some possible directions for future research.

Chapter 2

Background

In this chapter, we present the background of the DNN-based SE, which can be broadly divided into the following components. First, we give the mathematical expression of the noisy speech. Next, several speech acoustic features and their applications are explained. Then, we briefly discuss the DNN structures and hyper parameters. At last, we introduce several prevalent databases and metrics for objective evaluation of SE performances.

2.1 Noisy Speech Model

Although the interactions between background noise and clean speech are complicated, we aim to deal with the most common noise, that is, additive noise in our work. The time domain noisy speech can therefore be modelled as:

$$y\left(n\right) = s\left(n\right) + w\left(n\right) \tag{1}$$

where $s(n)$ is the clean speech, $w(n)$ the additive noise and $y(n)$ the noisy speech. n is the time index. Usually, $s(n)$ and $w(n)$ are assumed to be independent.

The noisy speech model in the T-F domain is obtained by STFT. It converts a time domain speech signal to a spectro-temporal spectrogram, where the harmonic structure of the speech can be observed clearly.

The STFT spectrogram of a clean speech signal $s(n)$ is defined as $S(k,l)$, with k and l indicating the frame index and frequency bin index of the STFT spectrogram, respectively. $S(k,l)$ has two expressions: one is in the rectangular coordinate system, which decomposes $S(k,l)$ into a real part and an imaginary part,

$$S\left(k,l\right)=\mathcal{R}\{S\left(k,l\right)\}+\mathcal{I}\{S\left(k,l\right)\} \tag{2}$$

the other is in the polar coordinate system, which decomposes $S(k,l)$ into a magnitude $\left|S\left(k,l\right)\right|$ and a phase $\phi_s\left(k,l\right)$.

$$S\left(k,l\right)=\left|S\left(k,l\right)\right|e^{j\phi_s(k,l)} \tag{3}$$

The corresponding spectrogram of the additive noise and the noisy speech can be donated as $W\left(k,l\right)$ and $Y\left(k,l\right)$. For simplicity, we denote the phase (or magnitude) of the clean speech and noisy speech as clean phase (or magnitude) and noisy phase (or magnitude) respectively.

2.2 Acoustic Features

Acoustic features, as the input of DNN, play an important role in supervised learning. When the acoustic features are able to fully and precisely represent the speech signal, the DNN-based system is more likely to obtain better enhancement performance, even without a powerful learning machine [68].

Early studies use single such as LPS, or a few features in SE [35]. The subsequent study includes exploring the performance with a combination of several features. In a more systematic literature research [69], an extensive list of 16 acoustic features are examined for SE at low SNRs. These features are extracted from different domains such as Mel-domain, modulation-domain, and gammatone-domain; and obtained with various techniques including linear-prediction, zero-crossing detection, autocorrelation, medium-time-filtering, and pitch analysis. We then introduce the most widely used acoustic features in detail below.

Amplitude modulation spectrum (AMS) [70] is recently used for speech segmentation in as a useful representation [71]. The detection of envelope fluctuations is a fundamental ability of the human auditory system which plays a major role in speech perception. Consequently, computational models have tried to exploit speech and noise specific characteristics of amplitude modulations by extracting so-called amplitude modulation spectrogram (AMS) features.

Relative spectral transform and perceptual linear prediction (RASTA-PLP) is another widely used feature in speech recognition [72, 73]. PLP is a popular representation in speech recognition [72]. The PLP technique derives an estimate of the auditory spectrum by using psychophysics of hearing, and is more consistent with human perception. RASTA filtering [73] is often coupled with PLP for robust speech recognition. Subsequently, PLP analysis is undertaken on this filtered spectrum. In fact, RASTA filtering serves as a modulation-frequency bandpass filter, which emphasizes the modulation frequency range most relevant to speech while discarding lower or higher modulation frequencies.

Mel-frequency cepstral coefficients (MFCC) collectively make up a Mel-frequency cepstrum. They are derived from a type of cepstral representation of the speech and first used in speech recognition [74]. In Mel-frequency cepstrum, the frequency bins are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bins used in the normal cepstrum. This frequency warping can allow for better representation of sound.

Relative autocorrelation sequence MFCC (RAS-MFCC) [75] is designed to improve the robustness of MFCC in the presence of additive noise by suppressing the noise in the autocorrelation domain. Firstly, the autocorrelation sequence is computed for each frame of an input signal. A high pass filter is then applied to the temporal trajectory of each dimension of autocorrelation sequences to suppress slow-varying components. The filtered autocorrelation sequences are treated as the input to the standard MFCC procedure to derive RAS-MFCC. Similar to RAS-MFCC, the autocorrelation sequence MFCC (AC-MFCC) [76] and phase autocorrelation MFCC (PAC-MFCC)

[77] are another two well-known features which aim to reduce the interference from background noise for MFCC.

The main idea of the gammatone frequency cepstral coefficients (GFCC) is based on an auditory periphery model that imitates the human cochlear filtering mechanism [78]. The auditory model is represented by a bank of Gammatone filters that decompose the input speech into a T-F representation. The Gammatone filters are obtained historically from several psychophysical and physiological observations of the human auditory periphery. The input signal is passed through the gammatone filters to obtain the subband signals, which are then decimated to 100 Hz. The GFCC is derived by applying cubic root compression and performing discrete cosine transform (DCT) to the magnitude of the decimated signals.

Power normalized cepstral coefficients (PNCC) is a recent feature for auto speech recognition that utilizes medium-time processing to mitigate noise corruption and employ power-law compression instead of log compression in traditional features [79]. Based on the medium-duration temporal analysis, the main idea behind PNCC is to subtract background noise by performing asymmetric filtering and temporal masking to the power spectrum of the input speech signal, where the power spectrum is integrated using gammatone frequency integration. Finally, the PNCC is obtained by applying power-law nonlinearity and DCT to the processed power spectrum.

Zero-crossing with peak-amplitudes (ZCPA) is also originated from speech recognition [80]. It is to detect the sign changes for every two adjacent samples of discrete signals. In general, an HF signal is assumed to have a large zero-crossing rate, while an LF signal has a small one. As such, the zero-crossing rate can be used to separate the background noise and desired speech when the noise and speech have different frequencies. To compute ZCPA, an input signal is first decomposed into subband signals by a 32-band gammatone filterbank. For each frame of the subband signal, the intervals between every two upward zero-crossings are calculated and then classified into 31 frequency bins. After adding a nonlinear-compressed peak amplitude to the corresponding frequency bin within each interval, the frequency bins are accumulated across all sub-bands and form a histogram, i.e. ZCPA.

Pitch-based features are first proposed in [81] for speech segmentation and are computed in T-F domain with pitch analysis. Based on the cochleagram of an input signal, the pitch-based features are derived for each T-F unit to capture how likely a unit is dominated by the target speech by utilizing periodicity and instantaneous frequency. The pitch-features are then used in supervised speech separation [82], where the ground truth pitch is used during training while the pitch is estimated by a recently proposed robust pitch tracker, PEFAC [83], is used during testing.

Multi-Resolution Cochleagram (MRCG) is a new acoustic feature proposed in [69], which encodes multi-resolution power distributions in the T-F representation of a signal. The authors combine four cochleagrams at different resolutions to construct the MRCG feature. A high resolution cochleagram captures the local information while three low resolution cochleagrams capture spectro-temporal contexts at different scales.

As there are various existing acoustic features, recent studies also investigate feature selections to obtain better enhancement performance. In [82], the acoustic features are selected by group Lasso method [84] and are examined the speech segmentation performance with different feature sets. Four features are then recommended as a complementary feature set comprising AMS, RASTA-PLP, MFCC and GFCC. Afterwards, the feature set is widely applied in SE. In [69], the authors compared the classification accuracy of each T-F unit with the estimated IBM, where the mask is estimated by DNN with different individual features as input. The results indicate that the gammatone-domain features such as MRCG and GFCC consistently outperform the other features. The authors also pointed out that the poor performance of pitch features is largely due to inaccurate estimation at low SNRs, as the ground-truth pitch is shown to be quite discriminative. Recently, a more comprehensive study considered the enhancement performance of speech denoising, separation and dereverberation [85]. The study also evaluates the performance of the DNN-based masking method [36], and reveals that the MRCG is the best under matched noises while PNCC is the best under unmatched noises in both anechoic and reverberant environment. For feature combination, this study concludes that the most effective feature set consists of PNCC, GFCC, and log-MFCC.

2.3 Basic Knowledge of DNN

Limited by the computing capability of the computer and the gradient vanishing problem in the training process, the DNN does not become popular until a breakthrough in DNN training was made by Hinton et al. [86], in which uses layer-wise unsupervised pretraining is employed to properly initialize a DNN before fine tuning. Afterwards, the DNN has been widely adopted in many applications including SE. Different architectures and hyper parameters settings are also investigated to improve the performance of the DNN.

2.3.1 Architectures

Although the DNN always consists of multiple layers between the input and output layers and each layer has the same components, such as neurons, weights, biases, and activation functions, the architecture of DNN varies, which makes different networks suitable for specific tasks.

FNN

An FNN denotes a conventional multilayer perception with many (often more than two) hidden layers [35]. Fig. 1 depicts an example process of using an FNN to learn the mapping between the noisy speech magnitude spectrum and the clean speech one. The FNN is fully-connected and has a total of five layers that include an input layer, three hidden layers and an output layer. The output of current layer serves as the input of the next layer and the non-linear mapping capability of the FNN relies on the activation function of each neuron in the hidden layers. The solid line reflects the forward propagation that obtains the estimated output, while the dashed line reflects the back propagation that adjusts the parameters of the FNN with the goal to minimize the prediction error, which is usually measured by a cost function between the estimated output and the desired output.

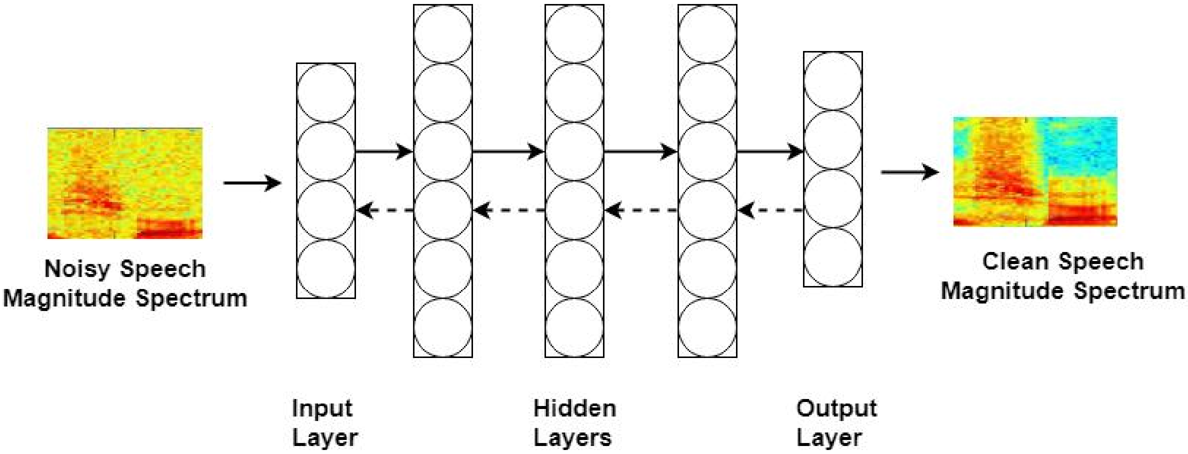


Figure 1: Architecture of FNN

CNN

The CNN is a special class of DNN with share-weights architecture and translation invariance characteristics, and can be used for speech processing [87]. Fig. 2 is an example of using a traditional CNN to enhance the noisy speech magnitude spectrum. The first layer of CNN is the convolution layer, which consists of a number of feature maps. The convolution operation in CNN means applying a filter (or kernel) on the input data to produce a feature map. Using different filters to perform multiple convolutions on an input results in distinct feature maps. The final output of the convolution layer is obtained by stacking all these feature maps together. A pooling layer is added on top of the convolution layer to compute a lower resolution representation of the convolution layer feature maps. The max pooling and mean pooling are two most widely used techniques. The convolution-pooling pairs can be stacked up to obtain higher level features. At last, the fully-connected layer is stacked for using the features for regression or classification tasks. It should be mentioned that the pooling layer and fully-connected layer are no longer necessities in recent new CNN structures. The former has the disadvantage of information loss and is replaced by dilated convolution [88], while the latter is of high-complexity with large parameters and can be substituted with another convolution layer [56].

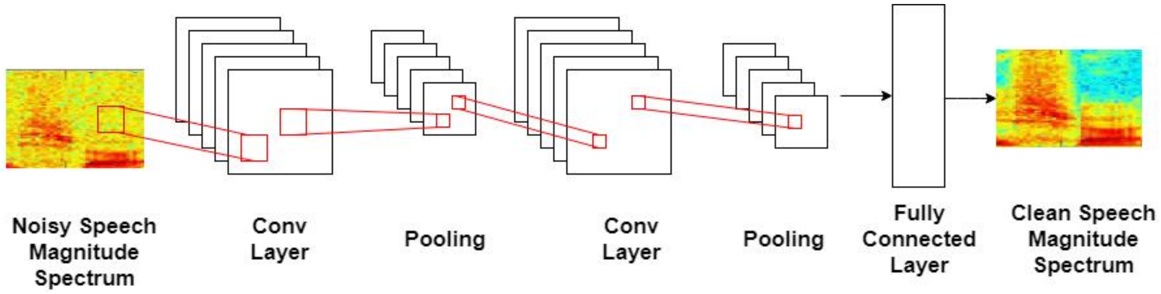

Noisy Speech Magnitude Spectrum
Conv Layer
Pooling
Conv Layer
Pooling
Fully Connected Layer
Clean Speech Magnitude Spectrum

Figure 2: Architecture of CNN

RNN

A recurrent neural network (RNN) is a class of neural network models where the term recurrent means that at least one feed-back connection is formed among its neurons, typically in the hidden layers. Such a recurrent connection is associated with the time-delay operation, which gives rise to the memory structure of RNN and potentially allows the RNN to model the sequential data. Unlike FNN, which processes each input sample independently, RNN treats input samples as a sequence and models the dynamics over the temporal dimension. This time dimension introduced by RNN is more flexible and infinitely extensible, while FNN does not share such characteristic no matter how deep the FNN is [89]. In other words, an RNN can be viewed as an FNN with infinite depth [90]. A speech signal is time-varying which exhibits strong temporal structure, and the current frame of the signal has dependencies with the previous frames based on linear prediction. Therefore, RNN is a natural choice for modelling the speech sequence and has been demonstrated to be well suited for speech processing [91]. The RNN training typically employs back-propagation through time [92], which is difficult to be well-trained due to the vanishing or exploding gradient problem [93]. To alleviate this problem, an RNN variant named LSTM is introduced [94].

In the LSTM network, the neuron in each hidden layer is replaced with a special unit called memory block, which has a self-connected memory cell to remember the temporal state and control the gradient and information flow. More specifically, a memory cell has three gates: input gate, forget gate and output gate. The forget gate controls how much previous information should be retained, and the input gate controls how much current information should be added to the memory cell. By opening and closing the gates, an LSTM allows relevant contextual information to be maintained in memory cells to improve RNN training. Fig. 3 shows the application of using an LSTM network for clean speech magnitude spectrum estimation. Unlike the FNN, whose information pathway is only from the input layer to the output layer, the dashed line in Fig. 3 represents temporal modeling where the information also passes along the time dimension.

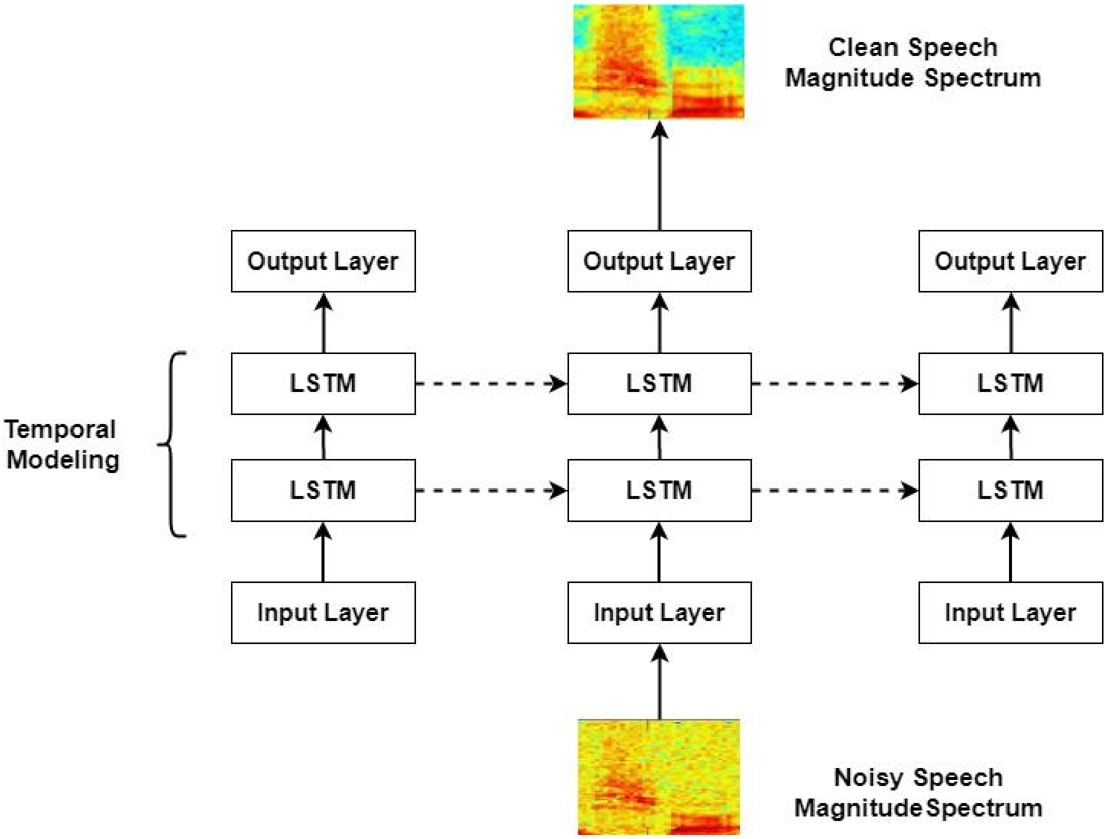


Figure 3: Architecture of LSTM

GAN

Despite the short history of GAN [95], this new network has attracted researcher's attention since published. GAN belongs to the generative model that learns to map samples **z** from some prior distribution $\mathcal{Z}$ to samples **x** from another distribution $\mathcal{X}$ of the real training data such as images or audio. The component within the GAN structure consists of two networks: the generator (G) and the discriminator (D). The main task of G is to imitate the real data distribution, that is to generate samples which are close to those of the training data. D typically works as a binary classifier and receives inputs from both real samples of the dataset and the generated samples in G. The goal of D is to classify the samples from the training dataset as real and those from G as fake. This framework is analogous to a two-player adversarial game that trains G to generate samples to fool D, while D learns to better tell the difference between the real data and generated data. Fig. 4 illustrates the case of employing GAN for enhancing the magnitude spectrum of the noisy speech [40]. G is structured similarly to an auto-encoder which takes the noisy magnitude as input and attempts to generate the enhanced speech. D classifies the enhanced speech as fake and the clean speech as real, and transmits the information to G and guides G to correct its output towards the realistic distribution. In the training stage, this adversarial learning drives both G and D to improve their accuracy until the generated enhanced speeches are indistinguishable from ideal clean speech. In the testing stage, only the trained G will be used for the task of SE.

2.3.2 Hyper Parameters

The performance of DNN-based method is not only determined by its architecture, but also the selection of the hyper parameters during the training process. As such, a short discussion of parameters is introduced as below. With proper settings, the DNN is more likely to obtain better results.

Neurons and Layers

The numbers of the hidden layers and the neurons at each layer are the primary parameters that should be considered. Firstly, the current layer works as an extractor to learn the features from the previous layer, which requires the number of neurons at each layer to be large enough to capture the essential information, especially for the lower layers. Actually, a bottleneck problem might happen if one of the layers has fewer neurons, which will significantly deteriorate the performance. Secondly, increasing the number of layers strengthens the learning capability of the DNN in theory.

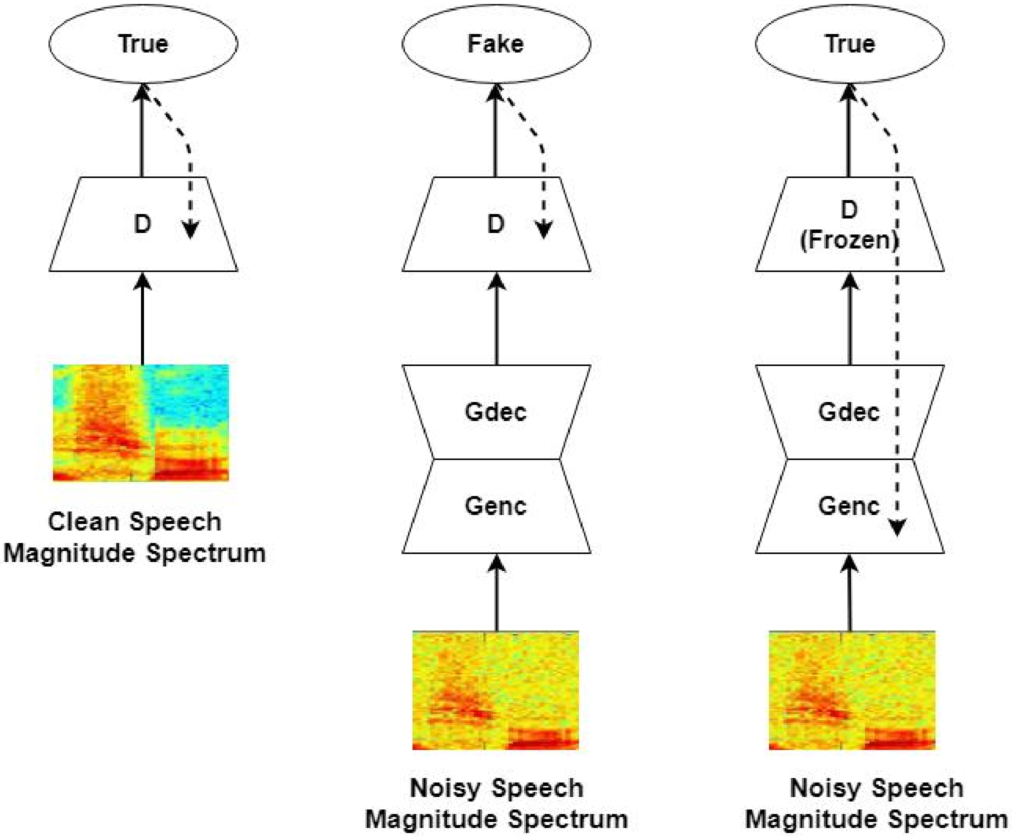


Figure 4: Architecture of GAN

However, a DNN with many hidden layers triggers the gradient vanish problem in training and is easier to overfit to the training data. In the SE task, the authors in [96] have found that the DNN achieves the best performance with 3 hidden layers and 2048 neurons in each hidden layer.

Activation Function

The activation functions are usually non-linear functions in each layer between the input feeding the current neuron and its output going to the next layer, which can help the network learn complex distribution, compute and model almost any function, and provide accurate predictions. Below, we introduce several common activation functions.

Linear function: Its equation takes the form $f(x) = cx$. Here, the activation is proportional to the input x, where c can be any constant value. As such, if the model only contains layers with linear functions, it is impossible to use back-propagation to train the model since the derivative of linear function is always a constant. In addition, no matter how many layers the model has, the final activation function of the last layer is nothing but just a linear function of the input of the first layer. Therefore, the linear function is only used in the output layer.

Sigmoid function: It is a well-known activation function whose mathematical definition is given as,

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

Sigmoid function has a smooth gradient, and its output is conveniently between 0 and 1 for all x. However, there are three major disadvantages in using the sigmoid function. First, if the value of x goes too large or too small, its gradient value approaches zero, which causes gradient vanishing problem. Second, the exponential function makes the sigmoid function computationally expensive. Finally, the output of sigmoid function is not symmetric around zero.

Hyperbolic tangent function: This function is also named tanh function and is actually mathematically shifted version of the sigmoid function,

$$\tanh(x) = \frac{2}{1 + e^{-2x}} = 2 * \mathrm{sigmoid}(2x) - 1 \tag{5}$$

Tanh function works better than sigmoid function because it is zero-centered, which makes it easier to model inputs that have strongly negative, neutral, and strongly positive values. However, tanh function still suffers the problem like sigmoid function, that is, gradient vanishing and computationally expensive.

ReLU: ReLU is the abbreviation of rectified linear unit, and is the most widely used activation function, which is given by $\mathrm{relu}(x) = \max(0, x)$. It gives an output x if x is positive and 0 otherwise. ReLu is less computationally expensive than tanh and sigmoid because it involves simpler mathematical operations. However, it has the dying ReLU problem, which means that when the input is negative, the gradient of ReLU becomes zero. Thus, the network cannot perform back-propagation. This can be solved by using the leaky ReLU function.

Batch Size

Neural networks are trained using back propagation algorithm, which involves updating the model weights and biases based on the gradient of the prediction error. This gradient is a statistical estimate with a batch of training samples. The choice of the batch size will affect both the convergence speed and the resulting model. Based on the different batch size, the gradient descent can be divided into the following categories:

- Batch Gradient Descent: Batch size is set to the total number of examples in the training dataset.
- Stochastic Gradient Descent: Batch size is set to one.
- Mini-batch Gradient Descent: Batch size is set to more than one and less than the total number of examples in the training dataset.

Theoretically, increasing the batch size helps adjust the weights and biases in a way that will improve the performance of the model. Given that the DNN training process usually involves large datasets, the batch size is rarely set to the size of the training dataset. The small batch size offers two main benefits including a regularizing effect and a lower generalization error. In SE task, the bath size varies from 128 to 2048 works well.

Learning Rate

Learning rate is a configurable parameter that controls the rate or speed of the model training during performing back-propagation. Specifically, it refers to the amount that the weights of the model are updated with each time they are updated. The values of learning rates often lie in the range between 0 and 1. Choosing the learning rate is challenging as a small value may result in a long training process that could get stuck, whereas a large value may result in learning a sub-optimal set of weights too fast or an unstable training process. Unfortunately, analytically calculating the optimal learning rate for a given model on a given dataset is a complex task. Instead, researchers usually discover a good learning rate through trial and error. Further, the batch size also affects the selection of learning rate. In general, smaller batch sizes are better suited to smaller learning rates.

Epoch

Epoch is defined as the times that the training algorithm will work through the entire training dataset. An epoch is comprised of one or more batches. In one epoch, each sample in the training dataset will be used to update the model parameters. The number of epoch is traditionally large, allowing the training algorithm to run enough times to minimize the loss function of the DNN. In practice, one could observe the curve that the value of loss function changes along with the number of epoch. This curve can help to diagnose whether the model is underfitting, overfitting, or suitably fits to the training dataset.

Momentum

It is well-known that the convergence speed can be improved when training a neural network with all the previous gradients for weights updating instead of only the current one [97]. Specifically, an exponentially weighted average of the prior gradients to the weight can be included when the weights are updated. In the DNN training, this is typically achieved with a simple technique named momentum, which controls the amount of the past gradients used in the updating. The application of momentum is able to smooth the optimization process and reduce the variance of the gradient estimation, which avoids the oscillation problems and speeds up the training. The value of momentum ranges from zero to one, where common values such as 0.9 and 0.99 are used in practice.

Dropout Rate

Although DNN is a powerful machine learning system, overfitting could be a serious problem due to DNN's large number of parameters. Dropout is a popular way to prevent overfitting, which is performed by randomly omitting a certain percentage of the neurons in each hidden layer for each epoch during training. In this case, the remaining neurons depend less on other neurons to learn the underlying patterns of the data. As such, the random dropout may break up the corporations of the neurons and reduce the learning capacity of DNN, but dropout improves the performance of DNN since this technique helps DNN generalize to unseen data. Typical values of dropout rate for hidden layers are in the range 0.5 to 0.8 [98].

2.4 Objective Evaluation

To date, researchers have put forward numerous SE techniques. The rapid deployment of SE increases the need for speech quality evaluation, which depends largely on end-user opinion of perceived speech quality. In this section, two key components of speech quality measurements have been introduced, that is, the databases used for enhancement and the metrics for evaluation.

2.4.1 Databases

A large training database is required in order to obtain better performance with deep learning technique. For SE, the databases are prepared from two aspects: clean speech and noise.

Clean Speech

The clean speech database should embrace good audio characteristics and text characteristics. Audio characteristics include the total duration of speech, speech quality, the number of speakers, and the attribute of speakers such as the distribution of gender and age group, and their dialect, accent, and tempos. Text characteristics refer to the language of speech and the richness of contexts, such as the number of words and the frequency a word appears. For DNN-based SE, the duration of speech, the number of speakers, and speech quality are the first three considerations. We introduce some widely used clean speech databases below.

1) *IEEE corpus* [99]: This corpus is recommended by IEEE subcommittee for speech quality measurements. The text contents of IEEE corpus, also known as Harvard sentences, which are

a collection of sample phrases that are originally used for standardized testing of communication systems such as Voice over IP, cellular, and other telephones. The corpus is then used for speech recognition and enhancement where standardized and repeatable sequences of speech are needed. The corpus consists of 72 lists with 10 utterances in each list. The utterances are produced by three male and three female speakers, and are phonetically balanced sentences that use specific phonemes at the same frequency they appear in English. The original sampling frequency of the utterance is 25 kHz.

2) *TIMIT* [100]: The TIMIT corpus is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers, including 192 female speakers and 438 male speakers, and the speakers cover eight major dialects of American English. Each speaker read 10 phonetically rich sentences, which results in 6300 utterances in total and a duration of approximately 5 hours. The TIMIT corpus includes time-aligned orthographic, phonetic, and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance.

3) *TED-LIUM* [101]: This corpus is a public database that is developed by the LIUM (Laboratoire Informatique de l'Universite du Maine) for automatic speech recognition, based on the TED (Technology, Entertainment, Design) Talks in English. The first released version [101] is built during the international workshop on spoken language translation campaign, and is composed of a total of 774 talks with an average duration of 9 minutes per talk, representing 118 hours of speech: 82 hours of male and 36 hours of female. The second [102] and third versions [103] continue to enrich the dataset by collecting more TED talks, which makes the corpus fairly rich in terms of languages, with $2.56m$ words covered in talks. However, the recorded utterance is noticeably reverberant and affected by the environment noise such as cough and applause, since the TED talks are recorded in conference halls that were full of audiences.

4) *MUSAN* [104]: The MUSAN corpus is originally designed for voice activity detection (VAD) and music/speech discrimination. The dataset consists of music from several genres, speech from twelve languages, and a wide assortment of technical and non-technical noises. For the

speech sub-dataset, the total duration of this portion is about 60 hours, which can be divided into two parts. The first part is the read speech from Librivox (20 hours and 21 minutes), which is a multi-language speech dataset with approximately 50% English and 50% eleven other languages. The content of each speech file is an entire chapter of a book read by one speaker. The second part is the speech from US government hearings, committees and debates (40 hours and 1 minute). These files have been obtained from the Internet archive and the Missouri channel senate archives. These recordings are entirely in English.

5) *LibriSpeech* [105]: LibriSpeech is originally prepared for auto speech recognition. The audio file is recorded by reading the audio-books LibriVox, and the total duration of all files is approximately 1000 hours. The recorded speech is with a sampling rate of 16 kHz and carefully segmented and aligned. The dataset is divided into several subsets: two training sets with 100 hours and 383.6 hours, respectively, one development set of 5.4 hours, and one test set of 5.1 hours. For development set and test set, 20 male and 20 female speakers are drawn at random and assigned, and each speaker read approximately eight minutes of speech. For each speaker in the training set, the amount of speech was limited to 25 minutes, in order to avoid major imbalances in per-speaker audio duration. The LibriSpeech only contains English [105], while the subsequent work extends LibriSpeech to multi-language [106].

Noise

In our research, the noisy speech is constructed by the addition of clean speech and noise. Therefore, the collection of noise database is another required step. The noise signal can be recorded either in a simulated environment or in a real-world place. Generally, a noise dataset should contain more types of noise in order to obtain a deep model with higher generalization capability. Below, we briefly introduce several prevalent noise datasets.

1) *Aurora-2* [107]: In order to represent the most probable application scenarios for telecommunication terminals, the noises in this corpus are recorded at different places: subway, babble, car, exhibition hall, restaurant, street, airport, train. Each noise sample in the corpus is 10 seconds

long, and is sampled at 8 kHz.

2) *NOISEX-92* [108]: This corpus is one of the earliest and most popular noise datasets, which contains white noise and a variety of non-stationary noises such as voice noise (babble), factory noise, HF radio channel noise, pink noise, and various military noises including fighter jets (Buccaneer, F16), destroyer noises (engine room, operations room), tank noise (Leopard, M109) and machine gun, and lastly, car noise (Volvo). Each noise signal has a duration of approximately 4 minutes. The dataset is widely used in speech processing tasks because of its rich types of noise, while a lack of real-world recordings is the biggest drawback of this dataset. Fig. 5 shows the spectrograms of 8 sample noises in NOISEX-92. As shown in the figure, the structure of non-stationary noises is more complex and irregular than that of white noise, thus to remove the non-stationary noise is more difficult, especially for the babble noise.

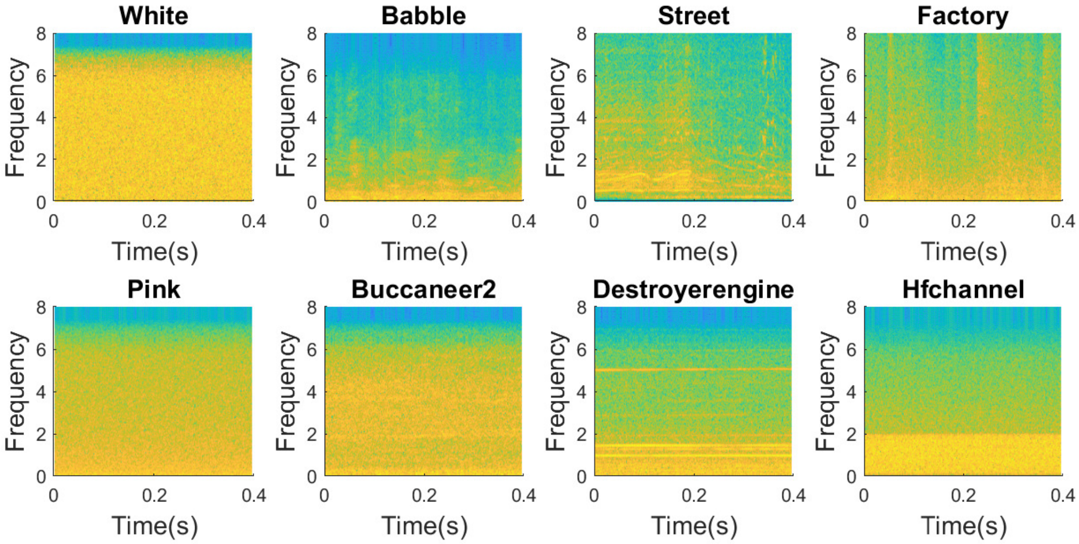


Figure 5: Spectrograms of different noises.

3) *CHiME-5* [109]: CHiME is a competition of speech separation and recognition which is held every two years. Each time the organizer will publish its corresponding dataset for competition. After publishing, the dataset is also widely used in academic research. For the CHiME-5, the background noise is real-life noise which is made up of the recording of twenty separate dinner parties that are taking place in real homes, with three different locations: kitchen, dining, and living

room. The recording in each location lasts at least half an hour. Four participants (two hosts and two guests) participate in the party, and the party members are familiar with each other and behave naturally.

2.4.2 Metrics

There are two main groups of methods to assess speech quality: subjective listening test and objective metrics. Subjective assessment is the most trustful method as it reveals the real feelings towards a speech signal. However, conducting subjective assessments requires large resources and experienced listeners, which is quite a time-consuming task. Therefore, various objective metrics have been proposed to assess the processed speech signal, which evaluates with knowledge from psychoacoustics, semantics, linguistics. The objective scores should have a high correlation with the subjective test results. Below, we introduce several common objective metrics.

SegSNR

The segmental SNR (SegSNR) [110] can be evaluated either in the time or frequency domain. The time domain measure is perhaps one of the simplest objective measures used to evaluate SE algorithms. To use SegSNR, the original signal $x(n)$ and processed signals are required to be time-aligned. The time domain segSNR is defined as,

$$\mathrm{SNR}_{seg} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \frac{\sum_{n=Nm}^{Nm+N-1} \|x(n)\|^2}{\sum_{n=Nm}^{Nm+N-1} \|\hat{x}(n) - x(n)\|^2} \tag{6}$$

where N and M are the frame size and the number of the frames, respectively.

SDR

Source-to-distortion ratio (SDR) is a widely-used measurement in speech separation and source enhancement [111], which is given by,

$$\text{SDR} = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{7}$$

where s_{target} is the estimated target source signal with low distortion, and e_{interf}, e_{noise}, and e_{artif} are the error terms caused by the interferences, noise, and artifacts, respectively.

PESQ

PESQ is proposed in the ITU-T recommendation P.862 [112]. Compared with the SegSNR and SDR, which measure the speech distortion physically, PESQ aims to model the subjective listening test results and characterize the speech quality as perceived by users. Firstly the original and processed signals are equalized to a standard listening level, then aligned in time to correct for time delays, and then processed through an auditory transform to obtain the loudness spectra. The difference between the loudness spectra of the processed signal and that of the original signal is computed and averaged over time and frequency to produce the prediction of subjective MOS. Although PESQ is an objective metric for evaluating speech quality, it also reflects faithfully the subjective score of the processed speech.

STOI

Besides speech quality, speech intelligibility is also an important index when evaluating a processed speech signal. Short-time objective intelligibility (STOI) is then put forward in recent years for objective assessment of the speech intelligibility [113]. It extracts short-time envelope vectors of the original and processed signals, to compute the average of the correlations across the envelope vectors, and the average correlation is then taken as the intelligibility score. Experiments demonstrate that the STOI score yields a high correlation with subjective intelligibility score.

Chapter 3

DNN-Augmented Basic Kalman Filter for Speech Enhancement

3.1 Introduction

In this chapter, we develop a DNN-augmented basic KF method to remove the background noise in time domain. KF was first applied to SE in [5], and remains of particular interest due to its several advantages: (1) ability to handle and process non-stationary signals; (2) absence of musical noise in the denoised speech given ideal parameters; (3) possibility of enhancing both the speech magnitude and phase.

As mentioned in Section 1.2, the enhancement performance of Kalman filtering is largely dependent on the estimation accuracy of the AR parameters. Ideally, the AR parameters of the clean speech can lead to the excellent performance of the KF [5], but they are not accessible in practice. Therefore, various estimation algorithms have been proposed to obtain the above parameters from the noisy speech, which can be divided into two categories: online estimation [114–117] and offline estimation [118, 119]. The former algorithms usually estimate and update the denoised speech and the model parameters in an iterative manner, while the latters require a training stage on a clean speech database to predict the parameters beforehand.

Recently, the authors in [120] make the first attempt to employ DNN for LPCs estimation in the area of extending monaural signal to stereo signal. More specifically, the monaural signal is assumed as the mid signal for the extended stereo signal. In addition, the residual signal is synthesized with the AR speech model, whose LPCs are estimated with the DNN by learning the mapping from the LPCs of mid signal to the LPCs of the residual signal. Inspired by this research, in this chapter, we apply DNN to estimate the LPCs of clean speech from those of noisy speech. In particular, we propose a DNN-augmented KF for SE, with an objective to improve the performance of traditional Kalman filtering by estimating the AR parameters with DNN.

This chapter is organized as follows. Section 3.2 describes the proposed DNN-augmented KF, including the detailed process of Kalman filtering for speech denoising as well as the DNN-based parameter estimation for constructing the KF. In Section 3.3, we propose an HF component restoration scheme to further improve the performance of Kalman filtering-based SE. Performances of the proposed three basic Kalman filtering-based methods are evaluated in Section 3.4 in terms of objective performance measures. Conclusions are drawn in Section 3.5.

3.2 DNN-Augmented Basic Kalman Filter

The overall block diagram of our proposed SE system with DNN-augmented basic KF is depicted in Fig. 6. It consists of two stages: the training stage and the enhancement stage. In the training stage, a DNN is trained to learn the mapping from the noisy line spectrum frequencies (LSFs) to the clean ones. In the enhancement stage, a KF with the DNN-based estimated parameters is applied to the noisy speech to obtain the enhanced speech. The main components of the SE system are introduced in the following subsections.

3.2.1 Basic Kalman Filter

As mentioned in Section 2.1, the noisy speech $y\left(n\right)$ is an additive mixture of the clean speech $s\left(n\right)$ and the background noise $w\left(n\right)$, where $n \in \mathbb{N}$ is the discrete time index. As usual, $w(n)$ is

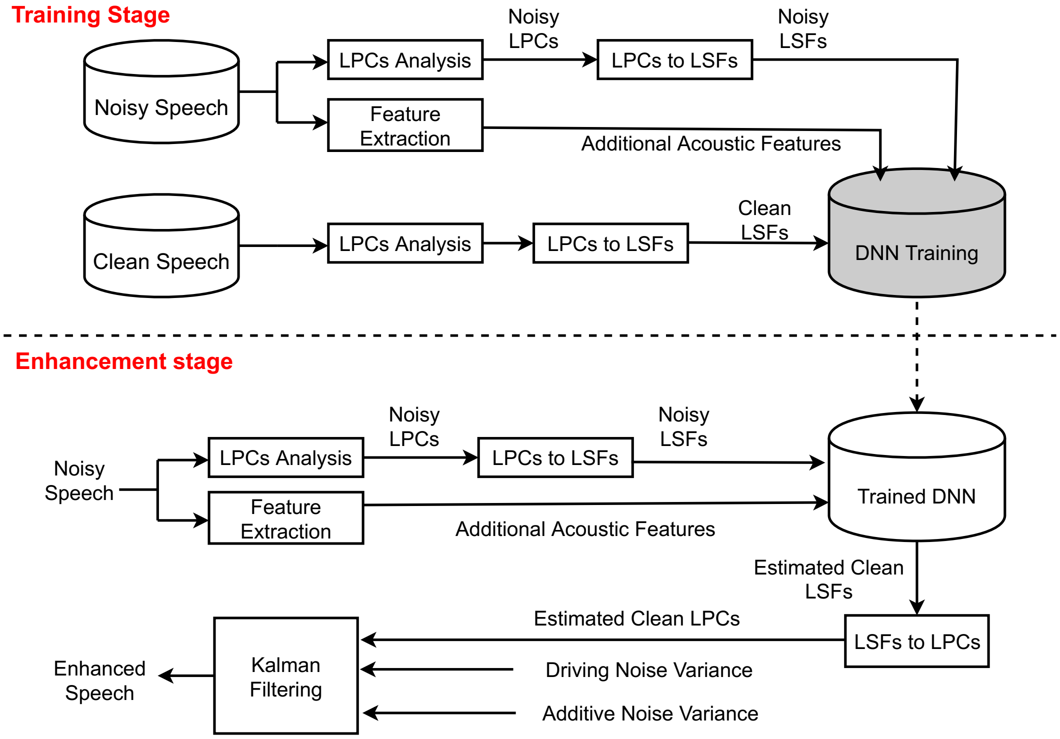


Figure 6: Block diagram of DNN-augmented basic KF for SE

regarded as a zero-mean white noise with variance σ_w^2, uncorrelated with $s(n)$. The clean speech $s\left(n\right)$ is usually represented by a linear model as a dynamic process of speech production. For the widely-adopted AR model, we have

$$s\left(n\right)=\sum_{i=1}^{p}a_{s,i}s\left(n-i\right)+v\left(n\right) \tag{8}$$

where $a_{s,i}$ are the LPCs of the clean speech, p the order of the model, and $v\left(n\right)$ the driving noise, i.e., a zero-mean white noise with variance σ_v^2.

To facilitate the KF presentation for SE, the above model equations for $s(n)$ and $y(n)$ can be rewritten in matrix form as,

$$\begin{cases} \mathbf{s}\left(n\right)=\mathbf{F}_s\mathbf{s}\left(n-1\right)+\mathbf{G}_s v\left(n\right) \\ y\left(n\right)=\mathbf{H}_s^T\mathbf{s}\left(n\right)+w\left(n\right) \end{cases} \tag{9}$$

where $\mathrm{s}\left(n\right)=\left[s\left(n-p+1\right),\ldots,s\left(n-1\right),s\left(n\right)\right]^T$ denotes the speech state vector. Moreover, the transition matrix $\mathbf{F}_s$ is given by

$$\mathbf{F}_s=\begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ a_{s,p} & a_{s,p-1} & \cdots & a_{s,2} & a_{s,1} \end{bmatrix} \tag{10}$$

and $\mathbf{H}_s=\mathbf{G}_s=\left[0,\cdots,0,1\right]^T\in\mathbb{R}^p$.

The denoising process with a KF amounts to recursively calculate an unbiased, linear MMSE estimate of the state vector $\mathrm{s}(n)$, given the corrupted speech $y(n)$. This process can be summarized by the following equations:

$$\begin{cases} e\left(n\right)=y\left(n\right)-\mathbf{H}_s^T\hat{\mathbf{s}}\left(n|n-1\right) \\ \mathbf{K}\left(n\right)=\mathbf{P}\left(n|n-1\right)\mathbf{H}_s\left(\sigma_w^2+\mathbf{H}_s^T\mathbf{P}\left(n|n-1\right)\mathbf{H}_s\right)^{-1} \\ \hat{\mathbf{s}}\left(n|n\right)=\hat{\mathbf{s}}\left(n|n-1\right)+\mathbf{K}\left(n\right)e\left(n\right) \\ \mathbf{P}\left(n|n\right)=\left(\mathbf{I}-\mathbf{K}\left(n\right)\mathbf{H}_s^T\right)\mathbf{P}\left(n|n-1\right) \\ \hat{\mathbf{s}}\left(n+1|n\right)=\mathbf{F}_s\hat{\mathbf{s}}\left(n|n\right) \\ \mathbf{P}\left(n+1|n\right)=\mathbf{F}_s\mathbf{P}\left(n|n\right)\mathbf{F}_s^T+\sigma_v^2\mathbf{G}_s\mathbf{G}_s^T \end{cases} \tag{11}$$

where $\hat{\mathrm{s}}\left(n|n-1\right)$ is the *a priori* estimate of the current state vector $\mathrm{s}\left(n\right)$, given observations up to a time index $n-1$, i.e., $y(1),...,y(n-1)$, $\mathbf{P}\left(n|n-1\right)$ the predicted state error correlation matrix of $\hat{\mathrm{s}}\left(n|n-1\right)$, $e\left(n\right)$ the innovation, $\mathbf{K}\left(n\right)$ the Kalman gain matrix, $\hat{\mathrm{s}}\left(n|n\right)$ the filtered estimate of state vector $\mathrm{s}\left(n\right)$, and $\mathbf{P}\left(n|n\right)$ the filtered state error covariance matrix of $\hat{\mathrm{s}}\left(n|n\right)$. The denoised speech $\hat{s}\left(n\right)$ is finally given by

$$\hat{s}\left(n\right)=\mathbf{G}_s^T\hat{\mathbf{s}}\left(n|n\right). \tag{12}$$

We note that several parameters appearing in the above equations should be estimated or calculated from the noisy observations in order to perform Kalman filtering. Those parameters include the driving noise variance σ_v^2, the additive noise variance σ_w^2, and the transition matrix $\mathbf{F}_s$, which contains the LPCs of the clean speech model.

3.2.2 DNN-Based LPCs Estimation

The LPCs of the clean speech AR model are estimated with a trained DNN, which consists of two steps: LPCs-to-LSFs conversion and DNN-based LSFs estimation.

LPCs-to-LSFs Conversion

When employing DNN for LPCs estimation, the LSFs, instead of LPCs, are usually adopted as the training target, since the former have a well-contained dynamic range of values, while the latter has a larger dynamic range. Therefore, we can maintain the stability of the training part more easily in the LSFs domain.

In the training part, LPCs are calculated using both noisy and clean speech databases, and then converted into LSFs for the DNN training. In the enhancement part, the estimated LSFs are converted to LPCs for Kalman filtering. The conversion process [121] is briefly summarized below.

A short segment of speech under the linear prediction analysis model is assumed to be generated as the output of finite impulse response filter $A(z)$. In order to define LSFs, the p-th order linear predictor $A(z)$ is decomposed into symmetrical and anti-symmetrical parts, represented by the polynomials $P(z)$ and $Q(z)$, respectively,

$$\begin{aligned} P\left(z\right) &= A\left(z\right) + z^{-(p+1)} A\left(z^{-1}\right) \\ Q\left(z\right) &= A\left(z\right) - z^{-(p+1)} A\left(z^{-1}\right). \end{aligned} \tag{13}$$

The LSFs ω_i are expressed as the zeroes (or roots) of $P(z)$ and $Q(z)$ in terms of the angular frequency.

The conversion from LSFs back to LPCs requires to obtain $A(z)$. Since $A(z)$ is expressed as the linear combination of $P(z)$ and $Q(z)$, i.e., $A(z) = 0.5[P(z) + Q(z)]$, we can easily construct $A(z)$ by using the ordered LSFs ω_i of $P(z)$ and $Q(z)$, i.e.:

$$\begin{aligned} P\left(z\right) &= \left(1 - z^{-1}\right) \prod_{i=2,4,\cdots,p} \left(1 - 2z^{-1}\cos\omega_i + z^{-2}\right) \\ Q\left(z\right) &= \left(1 + z^{-1}\right) \prod_{i=1,3,\cdots,p-1} \left(1 - 2z^{-1}\cos\omega_i + z^{-2}\right). \end{aligned} \tag{14}$$

DNN-Based LSFs Estimation

For supervised training, the DNN architecture adopted in our method is an FNN with many levels of non-linear units to represent a highly non-linear regression function that maps noisy LSFs to clean ones. Besides the noisy LSFs, we also investigate the use of other possible acoustic features in combination with the LSFs to form an extended input feature set, in order to better explore the relationship between the noisy LSFs and the clean LSFs, we would first like to. In [82], the following four feature types are shown to have good performance when acting as input to DNN. They are AMS, RASTA-PLP, MFCC and GFCC. Then, we will investigate the performance when these four features and their deltas are combined with LSFs as our input feature set. The total dimension of the combined input feature set is 258, i.e., (12+2×(15+31+13+64))

The input features are computed for each frame of the noisy speech, and represented as a row vector $\mathbf{f}(m)$ with m denoting the frame index. To make full use of the temporal information of the speech, it is common to incorporate the features of adjacent frames into a single extended feature vector. Hence, the extended feature vector centered at the m-th frame is constructed as $\tilde{\mathbf{f}}\left(m\right) = \left[\mathbf{f}\left(m - m_0\right), \cdots, \mathbf{f}\left(m\right), \cdots, \mathbf{f}\left(m + m_0\right)\right]$, where m_0 is the number of adjacent frames to be included on each side. The value of m_0 is set to 2 in our experiment. Note that all the different features are normalized to the range $[0, 1)$ in order to balance the training errors.

The structure of the FNN is depicted in Fig. 1. Our FNN is composed of one input layer, one output layer and three hidden layers with 1024 units in each layer. This structure has been verified to yield the best results in[120]. The rectified linear unit (ReLU) model is employed for the hidden layers, while the linear model is used for the output layer.

Back propagation with the MMSE-based cost function between the estimated clean LSFs and the reference clean LSFs is adopted to train the DNN. During the training, our DNN can automatically learn the complex mapping from noisy LSFs to clean LSFs given sufficient training samples. The well-trained DNN will be used in the enhancement stage to obtain estimated clean LSFs from the noisy LSFs.

3.2.3 Variance Estimation

The variance σ_w^2 of the additive noise $w(n)$ is usually estimated and updated during the unvoiced frames. The calculation involves a VAD procedure [122] to detect whether a given speech frame is voiced or unvoiced. In this method, three different features are used to determine a voiced frame. They are short-term energy, spectral flatness measure, and the most dominant frequency component of the speech frame spectrum. An audio frame is marked as a speech frame, if more than one of the feature values go over the precomputed threshold as proposed in [122].

The variance of the driving noise $v(n)$ can be then estimated as:

$$\begin{aligned} \sigma_v^2 &= \sigma_y^2 - \sigma_w^2 \\ &= \mathrm{E}\left[y^2(n)\right] - \mathbf{r}_y^T\,\mathbf{a}_y - \sigma_w^2 \end{aligned} \tag{15}$$

where $\mathbf{a}_y = [a_{y,1}, \cdots, a_{y,p}]^T$ is the LPC vector of the noisy speech, and $\mathbf{r}_\mathbf{y} = \mathrm{E}\left[\mathbf{y}\,(n)\,y\,(n)\right]$ the autocorrelation vector of the noisy speech $y(n)$ with its past p samples, represented by the vector $\mathbf{y}(n) = [y(n-1), \ldots, y(n-p)]^T$.

3.2.4 Summary of the Enhancement Stage

In conclusion, the DNN-augmented basic KF consists of an off-line training stage and an on-line enhancement stage. The former aims to learn the mapping between the noisy LSFs from clean LSFs with the proposed FNN, while the latter contains the following main steps:

- Computing the LPCs of the noisy speech and then converting the LPCs to LSFs.
- Combining the noisy LSFs with the acoustic features of the noisy speech as the input feature set.
- Estimating the LSFs of the clean speech from the input noisy feature set with the trained FNN.
- Updating the additive noise variance during non-speech frame and computing the driving noise variance with Eq. (15).
- Performing Kalman filtering with the estimated parameters to the noisy speech, and the final enhanced speech $\hat{s}(n)$ as given by (12).

3.3 High-Frequency Restoration for Kalman Filtering

Despite the performance gain from the KF based methods, it is found that the enhanced speech suffers from the loss or attenuation of its HF component. To address this problem, subband KFs have been investigated in [123, 124], wherein the noisy speech is decomposed into HF and LF components. The iterative KFs with different parameters are then applied into the HF subband and LF subband separately. Experimental results demonstrate that the subband KF algorithm outperforms the fullband counterpart. However, the HF component is still suppressed relative to the LF component. In other words, the desired speech in the HF subband is removed together with the noise when conducting Kalman filtering.

In this subsection, we propose an HF component restoration algorithm for KF based SE to further improve the performance. Inspired by the speech bandwidth expansion [125], the Kalman filtering denoised speech is first divided into HF and LF components. The LF component, which is considered to be of good quality, is then used to restore the HF component. At last, the enhanced speech is resynthesized by the LF component of the KF denoised speech and the recovered HF component, where the HF component restoration is accomplished with a DNN model.

The overall block diagram of our SE system with Kalman filtering and HF component restoration is depicted in Fig.7. The system is composed of the off-line training stage and the enhancement stage. In the training stage, an FNN is trained to learn the mapping from the log-magnitude of the LF component to that of HF component for clean speech. In the enhancement stage, the noisy speech is first processed by a KF to obtain denoised speech. Subband analysis is followed to decompose the denoised speech into HF and LF components. Then, the estimated HF component is recovered with the estimated magnitude predicted by the well-trained DNN and the phase from the denoised speech. Finally, the enhanced speech is obtained by a combination of the estimated HF component and the LF component of the denoised speech.

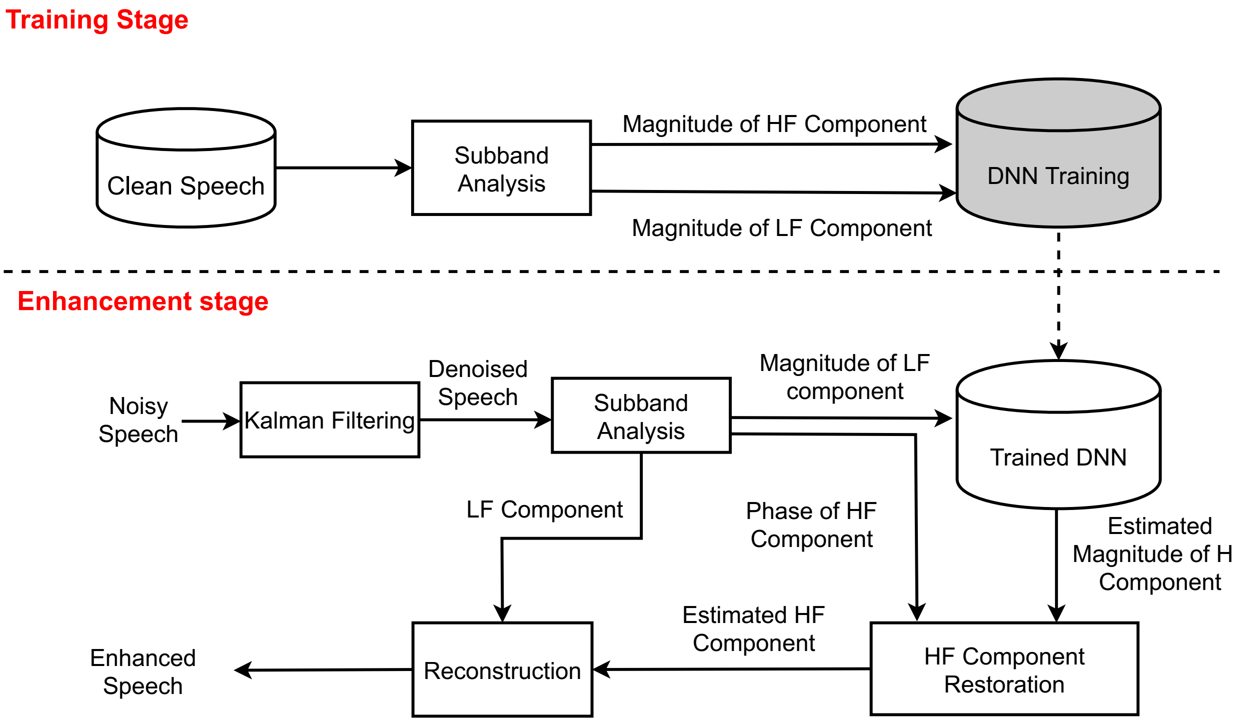


Figure 7: Block diagram of HF component restoration.

3.3.1 Training Stage

Data processing: the clean speech $s(n)$ is first transformed into its corresponding STFT spectrogram $S(k,l)$, with k and l denote the frequency bin and frame index, respectively. Subband analysis is then followed to divide $S(k, l)$ into HF component $S_H(n)$ and LF component $S_L(n)$ in T-F domain. For simplification, the index k and l will be omitted in the remaining discussion.

Feature and target: The magnitude of S_L is extracted as the input feature of DNN, while the magnitude of S_H is set as the training target. Since the magnitude spectrum usually has a very large dynamic range, the log-function and normalization are adopted to compress both the feature and target for better training. As usual, the features of the neighbouring frames are incorporated with the features of the current frame as an extended input feature set to make use of temporal information.

DNN structure: We adopted the same FNN structure as showed in Fig. 1 for the HF component estimation. The FNN has three hidden layers with 1024 units in each layer between the input layer and the output layer. The activation function used in the hidden layer is the ReLU, while a linear function is used in the output layer.

To update weights and biases until the network is able to achieve good performance, back propagation following a gradient-based optimization algorithm is commonly adopted. Back propagation computes the gradient, whereas stochastic gradient descent uses the gradients to train the FNN model, to minimize the value of the cost function, which is defined as the mean square error between the reference and the estimated log-magnitude spectrogram of the HF component

$$\mathrm{MSE} = \frac{1}{M} \sum_{m=1}^{M} \left[\left(ln|\hat{S}_H| - ln|S_H| \right)^2 \right] \tag{16}$$

where M denotes the speech frames, $|\hat{S}_H|$ the estimated magnitude and $|S_H|$ the reference one. The well-trained FNN will be used in the enhancement stage to obtain the estimated magnitude of the HF component from that of the LF component.

3.3.2 Enhancement Stage

The procedure of the enhancement stage can be briefly summarized as the following steps. Firstly, Kalman filtering introduced in Section 3.2 is applied to the noisy speech $y(n)$ for denosing in time domain. The denoised speech $d(n)$ is then transformed into T-F domain D. Subsequently, the D is decomposed into HF component $D_H(n)$ and LF component D_L by subband analysis.

Secondly, the aforementioned DNN-based HF component restoration algorithm is required to compensate the distortion in D_H. Here, the LF component D_L of the denoised speech is employed as input in restoration for the reason that the LF component is of high quality after Kalman filtering. The STFT spectrogram $\hat{D}_H$ of the recovered HF component is reconstructed with the estimated magnitude given by the well-trained DNN and the phase of D_H, i.e., $\hat{D}_H = |\hat{D}_H|e^{j\phi_{D_H}}$.

Finally, the STFT of the enhanced speech $\hat{S}$ is obtained by the subband synthesis of the restored HF component $\hat{D}_H$ and the unprocessed LF component D_L. The inverse STFT is performed to achieve the time domain enhanced speech $\hat{s}(n)$.

It should be pointed out that the Kalman filtering is performed in time domain, while the HF component restoration is accomplished in T-F domain. As such, the DNN-augmented Kalman filtering with HF component restoration as post-processing can be viewed as SE in both time and T-F domain.

3.4 Hybrid System of DNN-Based Speech Reconstruction and Kalman Filtering

In this section, we propose a two-level hybrid denoising system that exploits DNN-based speech reconstruction in conjunction with Kalman filtering in order to achieve improved performance. In the first level, a DNN is trained for the estimation of the speech magnitude spectrum, which is then used to reconstruct the clean speech. In the second level, another DNN is trained for predicting the LSFs of the clean speech, which will be transformed to LPCs. Meanwhile, the additive noise and driving noise variances are extracted from the reconstructed speech. Finally, a KF with the estimated parameters is applied to the reconstructed speech to obtain further enhancements. The main features of the hybrid system are summarized as follows.

- As well-known, the current deep learning based methods often suffer from performance degradation due to the data mismatch between the training and testing stages. Consequently, the reconstructed speech from DNN-based method inevitably contains residual noise in unmatched acoustic environment. By incorporating and combining Kalman filtering with a DNN-based speech reconstruction method, the hybrid system makes it possible to further reduce the residual noise in unmatched conditions.

- Further advantages of employing DNN include the following: First, DNN is used to estimate clean speech amplitude in order to perform preliminary SE. The additive noise and driving noise variances required for Kalman filtering are then more accurately estimated from the DNN pre-enhanced speech. Second, DNN is used to obtain accurate LPCs estimates which is critical for improved Kalman filtering.

- The speech reconstruction is performed in the frequency domain, that is, the reconstructed speech is obtained by synthesizing the estimated magnitude and the noisy phase spectra, while the denoising process of Kalman filtering is realized in the time domain. With such a combination, our hybrid system can be viewed as a joint estimator for both magnitude and phase of the spectra of the clean speech.

The overall block diagram of our hybrid system is depicted in Fig. 8. It consists of two stages: training stage and enhancement stage. In the training stage, we first extract noisy speech acoustic features, and then input them to two FNNs which are trained separately to learn the mapping from the noisy features to different targets: the magnitudes and LSFs of the clean speech. In the enhancement stage, the noisy speech features are extracted and processed by the well-trained FNNs to predict the clean magnitudes and LSFs. The estimated spectral magnitudes together with the noisy phase spectrum are then synthesized to obtain the reconstructed speech. Finally, a KF with the estimated parameters is applied to the reconstructed speech to obtain the enhanced speech. The key components and processing steps involved in the hybrid system are described in further detail below.

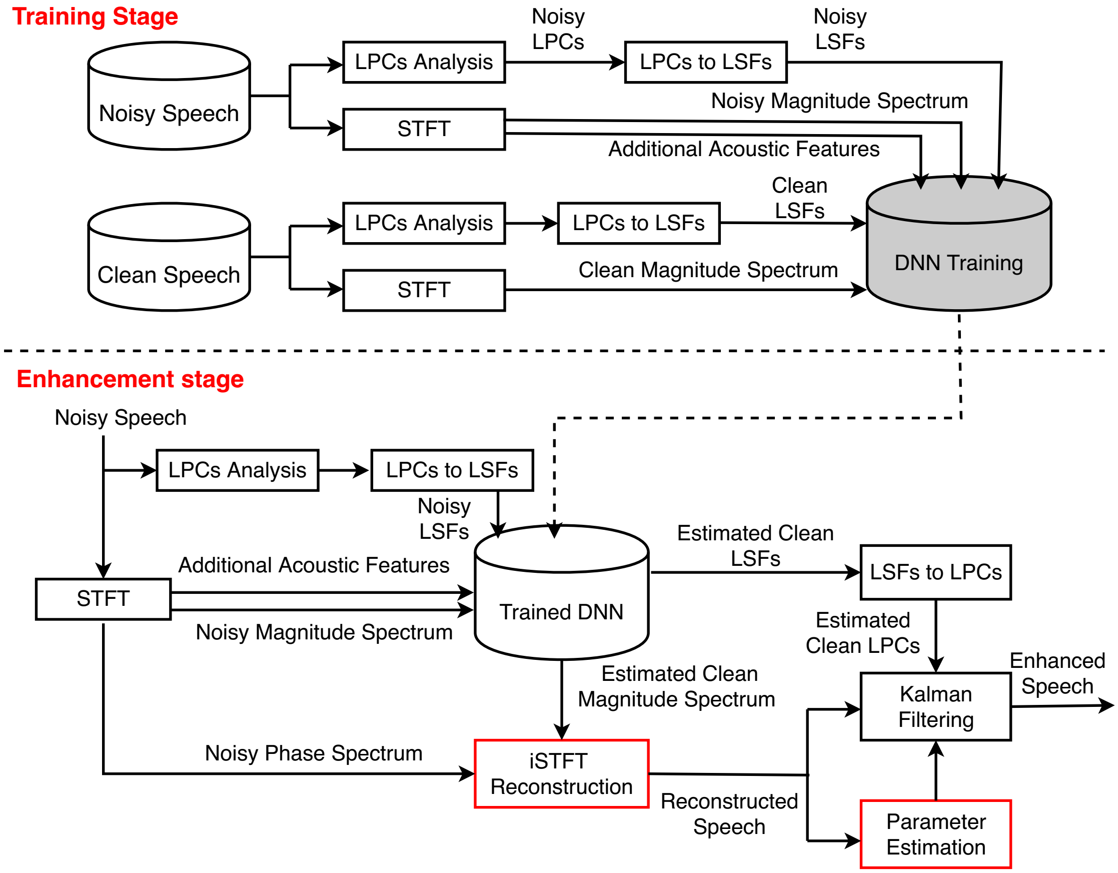


Figure 8: Block diagram of proposed hybrid SE system.

3.4.1 DNN Training

Two different training targets are set as the output of the DNNs, i.e., : the spectral magnitudes and LSFs. The magnitudes are employed in the speech reconstruction, while the LSFs are converted to LPCs as key parameters for Kalman filtering. The magnitudes are chosen as training target because the authors in [38] point out that, directly using magnitudes as training target can yield good performance and furthermore requires lower computational complexity. The LSFs are chosen as mentioned in Section 3.2.2.

For the input features, we adopt the acoustic features in [82] as additional input features in our work. Note that we include the LSFs into the input feature set when the training targets are LSFs, and included the speech spectral magnitudes of the speech spectrum when the targets are magnitudes. With these two specific feature sets, we are able to better learn the mapping from the noisy features to the training targets.

As shown in Fig.9, we employ two FNNs to estimate the spectral magnitudes and LSFs separately in our work. Employing two separate DNNs can provide better results than training only one DNN to learn the mapping to these two targets by multi-objective learning, because the LSFs and the magnitudes share little similarity in their structure. The input features of both FNNs are computed for each frame of the signal. To make full use of the temporal information of speech, it is common to incorporate features of adjacent time frames into a single feature vector. Moreover, we normalize different features into the range $[0, 1)$ in order to balance the training errors.

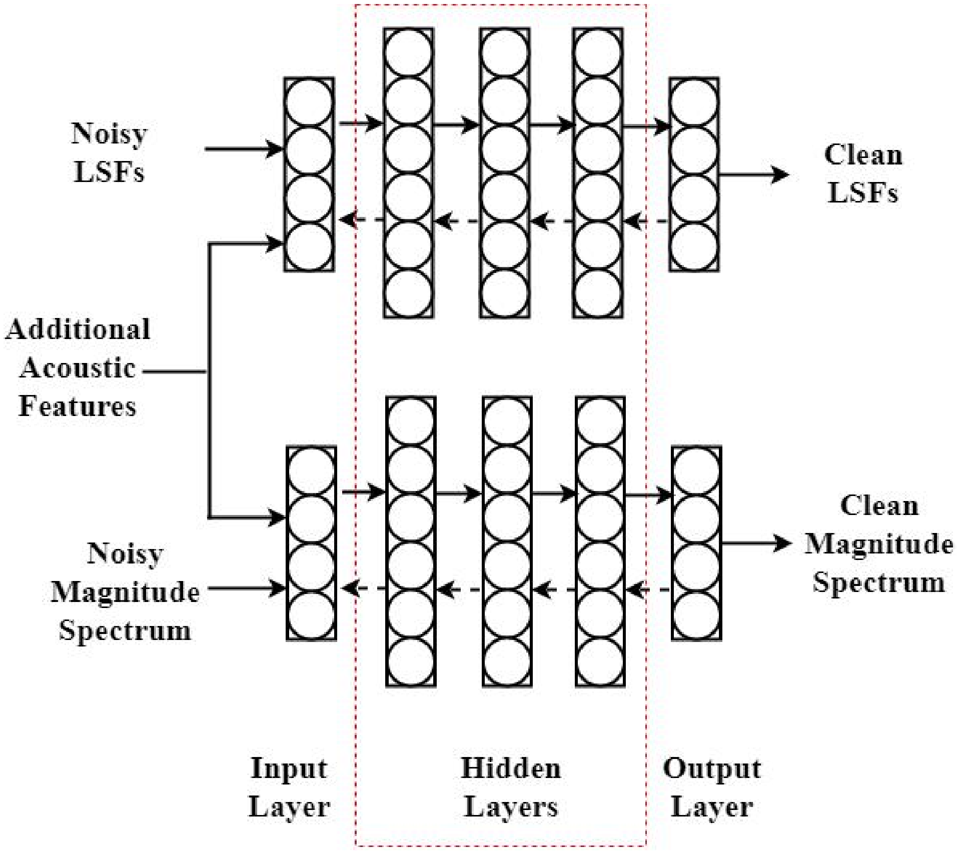


Figure 9: DNN structure in proposed hybrid SE system.

Although the targets are different, the settings for each FNN are the same in our work. Each network consists of an input layer, an output layer and three hidden layers, each comprising 1024 units. The linear activation functions are used in the output layer, whereas the rectified linear functions are used in the hidden layers.

Back propagation is used to adjust the weights and biases in the training part. The cost function for each training utterance is defined as the mean square error (MSE) of the magnitudes (or LSFs). The respective MSE is computed between the clean and estimated targets, i.e.,

$$\mathrm{MSE}_{\mathrm{MAG}} = \frac{1}{KF}\sum_{k=1}^{K}\sum_{l=1}^{F}\left(|\hat{S}\left(k,l\right)| - |S\left(k,l\right)|\right)^2 \tag{17}$$

or

$$\mathrm{MSE}_{\mathrm{LSF}} = \frac{1}{Kp}\sum_{k=1}^{K}\sum_{i=1}^{p}\left(\hat{\omega}\left(k,i\right) - \omega\left(k,i\right)\right)^2 \tag{18}$$

where $|S(k,l)|$ and $|\hat{S}(k,l)|$ are the clean and estimated magnitudes, respectively, with K indicating the number of frames and F the number of frequency bins, while $\omega(k,i)$ and $\hat{\omega}(k,i)$ are the clean and estimated LSFs, respectively, with i indicating the order index and p the AR speech model order.

3.4.2 Two-level Enhancement

The first level is speech reconstruction, which aims to obtain the STFT of the reconstructed speech, $R\left(k,l\right)$ by combining the estimated clean magnitudes from the well-trained DNN together with the noisy spectral phase values ϕ_y , i.e., $R\left(k,l\right) = \left|\hat{S}\left(k,l\right)\right| e^{j\phi_y(k,l)}$. The reconstructed speech $r\left(n\right)$ is then obtained by computing the inverse STFT of $R\left(k,l\right)$.

The second level is Kalman filtering, which further removes the residual noise in the reconstructed speech. Note that although we have used the noisy phases for synthesis in the first level, the reconstructed speech will be Kalman filtered in the time domain, which can be regarded as a joint form of enhancement of the magnitude and phase spectra.

The parameters of the KF are obtained as follow. Firstly, the LSFs are estimated from the FNN and then converted to LPCs to form the transition matrix $\mathbf{F}$. Secondly, the additive noise variance can be estimated during the speech-absent frames. Thus, estimation accuracy of σ_w^2 is highly dependent on the ability to detect the voice and unvoiced parts of the noisy speech. Here, the voice activity detector (VAD) algorithm [122] based on speech energy and spectral flatness is adopted for this purpose. Fig.10 depicts the VAD results of one noisy speech and its corresponding reconstructed speech. The blue waveform is the original clean speech. The decision line represents an unvoiced part when its value equals to 0, and a voiced part otherwise. The noisy speech is corrupted with pink noise at -3 dB. As seen in Fig.10, applying VAD to the reconstructed speech $r(n)$ rather than the noisy speech, helps make a correct decision of the unvoiced parts as seen in Fig.10, and in turn, improve the estimation accuracy of the additive noise variance.

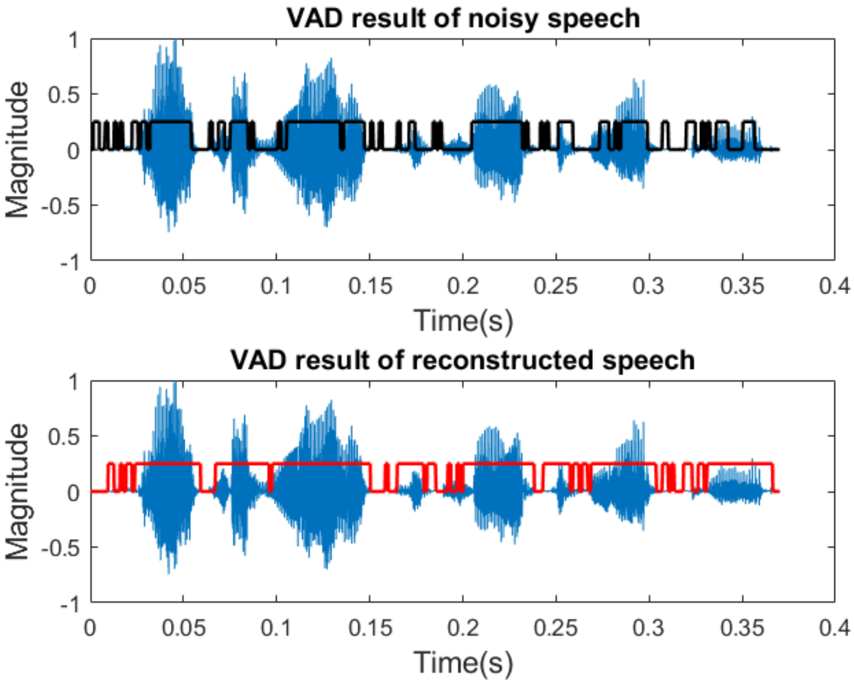


Figure 10: The VAD results of noisy and reconstructed speech.

For the estimation of the variance of the driving noise σ_v^2, we solve the Yule-Walker equations for the linear prediction model of the reconstructed speech, instead of using the estimation algorithm given in Eq. (15). Fig.11 shows the comparison of the estimated variance σ_v^2 of a noisy speech, which is corrupted with pink noise at -3 dB. Our algorithm (black) is closer to the true one (blue), which shows that the new algorithm achieves a better performance.

3.4.3 Summary of Hybrid System

The main processing steps of the proposed hybrid system are summarized as follows:

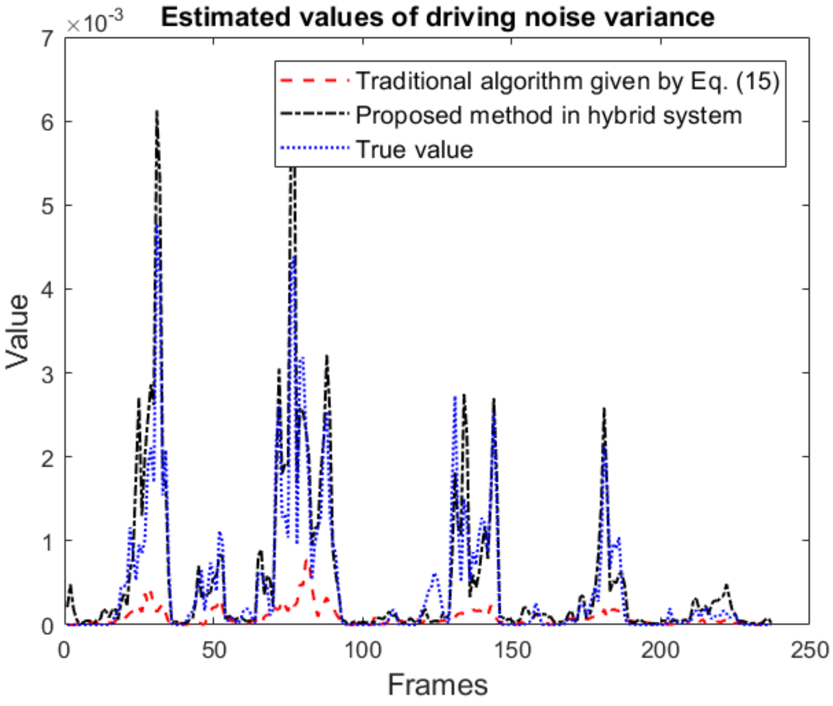


Figure 11: The estimation of driving noise variance through different methods.

1) Estimating clean LSFs and magnitudes from noisy features with the proposed DNNs.

2) Synthesising the reconstructed speech $r(n)$ with the estimated magnitude and the noisy phase spectra.

3) Converting LSFs to LPCs to form the state transition matrix.

4) Computing the the additive noise variance σ_w^2 and the driving noise variance σ_w^2 from the reconstructed speech.

5) Performing Kalman filtering to the reconstructed speech to obtain $\hat{\mathbf{u}}\,(n|n)$, and the final enhanced speech $d(n)$ as given by (12).

3.5 Experimental Results

3.5.1 Experimental Setup

The clean speech is selected from the IEEE sentence database[1] [99]. We choose 670 utterances for the training part and the remaining 50 utterances for the enhancement part. The noises are selected from the NOISEX-92 database [108], Four types of noises (babble, white, street and factory) are regarded as seen noise, and another four types (pink, buccaneer2, destroyerengine and hfchannel) as unseen noise.

For SE. In the training stage, the noisy speech is obtained by mixing clean training utterances with seen noise at four different levels of SNRs, i.e., -3dB, 0dB, 3dB and 6dB, which results in 10720 utterances. In the enhancement stage, both seen and unseen noises are mixed with clean testing utterances at the above mentioned four SNR levels. The number of noisy utterances used in the enhancement part is 800 for both seen and unseen noises. The sampling frequency for the speech and noise signals is set to 16kHz.

For HF component restoration, the deep model is trained only on clean speech database to explore the relationship between its LF and HF components. Since 670 utterances are not enough for deep learning, we repeat them for 16 times to get 10720 utterances.

Reference methods

To evaluate the performance of the proposed new system, we choose several existing approaches for comparison, which include one traditional Kalman filtering algorithm: Iter-KF; and three recent DNN-based methods, i.e.: FNN-MAG, FNN-IRM, FSEGAN. These are introduced briefly in the following.

Iter-KF [115]: The enhanced speech is obtained by iteratively performing conventional Kalman filtering, in which the LPCs are updated in each iteration.

[1] Available at website `https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip`

FNN-MAG [96]: An FNN is employed to directly explore the mapping from the noisy magnitude spectrum to the clean one. The enhanced speech is synthesized with the estimated magnitude and noisy phase spectra.

FNN-IRM [126]: An FNN is trained for better predicting the IRM. The estimated IRM is then applied to the noisy magnitude spectrogram to reduce the noise part, and the enhanced speech is then reconstructed from the masked magnitude and noisy phase spectra.

FSEGAN [40]. A least-square GAN is utilized to generate the clean speech magnitude spectrogram from the noisy one. The enhanced speech is reconstructed from the generated clean magnitude and noisy phase spectra.

Our proposed methods with basic KF can be concluded as below.

FNN-KF: An FNN is used to predict the LSFs for Kalman filtering. The noisy speech is processed by the KF to obtain the enhanced speech. We note that the FNN-MAG and FNN-IRM are frequency-domain SE methods, while the FNN-KF is a time domain method.

FNN-KFBE: The FNN-KF is adopted to obtain the Kalman filtered speech, whose HF component is then compensated with the FNN-based bandwidth extension algorithm.

Hybrid: A Hybrid system which consists of speech reconstruction and Kalman filtering. The reconstructed speech is first synthesizing with the estimated magnitude spectrum and the noisy phase spectrum, and is then further denoised with the FNN-augmented Kalmen filter.

In order to fairly evaluate the performance of the method proposed in this paper, we use the same DNN configuration in all the methods except FSEGAN. For FSEGAN, we adopt the settings provided in [40] and adjust other network parameters to optimize performance. For the remaining DNN-based methods, we use the standard FNN configuration. For FNN-MAG, FNN-IRM, and the HF component restoration, the Hamming window is selected to divide each utterance into 20 ms frames with a 10 ms frame shift (50% overlap). A 320-point DFT is then computed for each frame resulting in 161 samples. For FNN-KF, a rectangular window is used to divide the audio signals into 20 ms frames with no overlap. For the hybrid system, the STFT setting used in the magnitude spectrogram computation is the same as that of FNN-MAG, and the framing process in the LSFs estimation is the same as that of FNN-KF. In the implementation of the Kalman filtering algorithm, we set $\mathbf{u}(0|0) = \mathbf{0}$, $\mathbf{P}(0|0)$ as an identity matrix, and the speech AR order as $p = 12$.

For the DNN training, we adopt the same settings of hyper parameters for all tested methods, which is describes as following. The gradient descent optimization algorithm is Adam (Adaptive moment estimation). The batch size is 1024 and the training epoch is 20. The learning rate is linearly decreasing from 0.08 to 0.001 during the training process. The initial and finial momentums are 0.5 and 0.9, respectively. The drop out ratio is 0.2 in the hidden layers.

Objective metrics

To evaluate the enhancement performance, two objective metrics are selected: the PESQ measure and the STOI measure. PESQ and STOI evaluate the processed speech from two different aspects: speech quality and intelligibility, and a large objective score refers to a better performance for both metrics.

3.5.2 Investigation of Input Feature Set

In the training stage, we use the following feature sets as the input of our proposed system: LPS-only set, LSF-only set, multi-feature set consisting of AMS+RASTAPLP+MFCC+GFCC, and joint set formed by combining the LSF-only set with the multi-feature set. In this experiment, we investigate the performance of the proposed FNN-KF with these different feature sets when using FNN for LSFs estimation. The objective results of the enhanced speech are shown in Table 1.

The final enhanced speech for the LPS-only and LSF-only feature sets exhibit similar PESQ and STOI scores, while the dimension of LSF-only set is much smaller than the dimension of LPS-only set. In addition, the objective scores could be improved notably for the multi-feature and joint sets, which indicates that using more acoustic features provides useful additional information about the speech. Finally, the enhanced speech from the joint set achieves the highest PESQ and STOI scores. As a result, the joint set is considered as the optimal input feature set for the proposed methods considering both the dimension and the performance of all feature sets.

Table 1: Objective results with different feature sets in FNN-KF system

		-3dB	0dB	3dB	6dB
PESQ	Noisy	1.41	1.52	1.68	1.86
	LPS-only	1.61	1.84	2.02	2.19
	LSF-only	1.62	1.85	2.04	2.21
	Multi Set	1.67	1.89	2.10	2.27
	Joint Set	**1.71**	**1.93**	**2.13**	**2.30**
STOI	Noisy	0.66	0.72	0.78	0.83
	LPS-only	0.68	0.74	0.79	0.83
	LSF-only	0.67	0.73	0.79	0.83
	Multi Set	0.70	0.76	0.80	0.84
	Joint Set	**0.71**	**0.77**	**0.81**	**0.85**

3.5.3 Evaluation of Proposed Methods

Tables 2 and 3 show the average objective scores of the different SE algorithms on both seen and unseen noises respectively. In general, for the seen noise, the overall objective scores achieved by FNN-IRM and the proposed hybrid method are close, and are superior to the remaining methods. For the unseen noise, the overall objective scores clearly show that the proposed hybrid method performs better than the other DNN-based methods in most cases, except for the STOI score of FNN-IRM at 6dB SNR. A more detailed analysis of the results is provided in the following.

Seen noise

In the case of seen noise (Table 2), we serially present the evaluations of the proposed three methods, i.e., FNN-KF, FNN-KFBE and the hybrid system.

Evaluation of FNN-KF: We first compare our proposed FNN-KF with the traditional KF based methods. Obviously, Iter-KF achieves the worst performance among all tested methods, which is mainly caused by the inaccurate estimation of the AR parameters. The objective score can be significantly improved with the proposed FNN-KF, which infers that the adoption of FNN for LSFs estimation provides more accurate LPCs to further improves the enhancement capability of the KF. Moreover, compared with the DNN-based approaches, we note that the performance of FNN-KF is better than that of FSEGAN on our tested database. One possible reason could be that the generative model requires a larger amount of training data to learn the underlying distribution of the target features; otherwise mode collapse may happen in the training stage [127]. On the other side, the performance of FNN-KF is not as good as those of FNN-MAG and FNN-IRM. The performance degradation lies in the HF component distortion of the Kalman filtered speech, due to the inaccurate of the calculation of the two parameters in Kalman filtering, i.e., driving noise variance and additive noise variance.

Evaluation of FNN-KFBE: It is shown in Table 2 that the enhanced speeches from FNN-KFBE have better PESQ and STOI scores in comparison to those from FNN-KF, which demonstrates the advantage of introducing the HF component restoration as a post-processing for FNN-KF. In addition, by comparing the results between the enhanced speeches from the FNN-KFBE with respect to input SNRs, it can be found that the improvement is greater at high SNRs. One possible reason for this phenomenon is that the quality of the denoised speech at high SNRs is better than the one at low SNRs, which is beneficial to the restoration of the HF component. However, the bandwidth extension in HF component restoration algorithm uses the LF component of the clean speech to restore the corresponding HF component in training stage, while the LF component of the Kalman filtered speech is used in enhancement stage. The residual noise in the LF component leads to a degraded the restored HF component. Therefore, the performance of the FNN-KF is still not comparable to those of FNN-MAG and FNN-IRM.

Evaluation of the hybrid system: We compare the performance of the hybrid system with the other four FNN-related approaches, i.e., FNN-MAG, FNN-IRM, FNN-KF and FNN-KFBE. Obviously, FNN-IRM shows the best overall performance, especially in the high SNR region for PESQ. One possible reason for this outcome under matched condition is the use of different targets: for FNN-MAG and FNN-KF, the targets (clean magnitudes or LSFs) are the same across different noises and SNRs, and thus the FNN has to learn a many-to-one mapping; whereas for FNN-IRM, the targets (IRM) depend on the noise type and SNR, and thus the FNN is faced with the simpler task of learning a one-to-one mapping [126].

The hybrid system also exhibits better performance than FNN-KF and FNN-MAG. For FNN-KF, which has a limitation in accurately estimating the parameters σ_v^2 and σ_w^2 from the noisy speech, suffers distortion in the final output speech. For FNN-MAG, the quality of the enhanced speech is hindered by the residual noise, especially at lower SNR. The hybrid system, which can be regarded as a combination of FNN-MAG and FNN-KF, leads to a better enhanced speech because it employs the reconstructed speech as the input of Kalman filtering, and thus can provide more accurate estimates of σ_v^2 and σ_w^2, which in turn helps the KF better reduce the residual noises in the reconstructed speech. In addition, although the hybrid system and FNN-KFBE can be viewed as a two-level SE in both time and T-F domain, the performance of the hybrid system is much better because the hybrid system gains better denoised speech in the first level.

Compared to FNN-IRM, the hybrid system achieves about the same level of performance. The PESQ score is slightly better than FNN-IRM at -3dB SNR and a little worse at higher SNR, while the STOI scores for both methods are quite close at all SNRs. Hence, in the case of seen noise, our proposed hybrid system and FNN-IRM achieve the best performance among all the evaluated methods.

Table 2: Objective scores of different methods on seen noise

	PESQ				**STOI**			
Methods	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB
Noisy	1.41	1.52	1.68	1.86	0.66	0.72	0.78	0.83
Iter-KF	1.55	1.79	2.01	2.25	0.66	0.72	0.79	0.84
FNN-MAG	1.89	2.13	2.34	2.55	0.75	0.82	0.86	0.88
FNN-IRM	2.01	**2.28**	**2.47**	**2.67**	**0.80**	**0.84**	**0.88**	**0.91**
FSEGAN	1.85	2.02	2.19	2.35	0.70	0.75	0.80	0.84
FNN-KF	1.71	1.93	2.13	2.30	0.71	0.77	0.81	0.85
FNN-KFBE	1.75	2.02	2.20	2.39	0.73	0.79	0.84	0.88
Hybrid	**2.05**	2.23	2.44	2.61	0.79	**0.84**	**0.88**	0.90

Unseen noise

We first investigate the generalization capability of the tested methods by considering unseen noise. Upon comparison of the results in Table 2 and 3, we note that all the methods suffer from a performance degradation. Comparing the results in Table 2 and 3, we find that at high SNR, the performance of Iter-KF remains at a similar level as it belongs the class of unsupervised methods. In contrast, the objective scores of FSEGAN, FNN-IRM and FNN-MAG suffer a noticeable decrease, suggesting that the trained DNNs cannot achieve the same prediction accuracy under unseen noise. However, such a decrease in objective scores is not observed with FNN-KF and FNN-KFBE, whose PESQ scores now exceed those of FNN-MAG for SNR $\geq$ 0dB. This may be explained by the fact that the use of FNN in FNN-KF is limited to the LSFs estimation, while the core processing function, i.e. Kalman filtering, is a conventional method and therefore its performance should remain at a similar level whether in seen or unseen noise situations. While the performance of our proposed system drops slightly in the case of unseen noise, this degradation is not as significant as that observed with the FNN-MAG and FNN-IRM methods.

The overall performance of the proposed hybrid system is significantly better than the other methods in terms of both PESQ and STOI scores, except for the STOI scores of FNN-IRM at high SNRs. However, at high SNR, intelligibility is less of a concern, as it is not difficult to understand the speech in this case, while the speech quality remains our major concern, which is well handled by the proposed system as reflected by PESQ scores. At low SNR, the speech intelligibility is severely impacted by the additive noise and should be our priority task. Clearly, the proposed hybrid method gives better STOI scores in low SNR situations. In conclusion, the proposed hybrid system achieves the best overall performance in unseen noise, after considering the various aspects of objective evaluation metrics.

We also characterize the enhancement performances of the proposed hybrid system on the different types of noise. The objective scores of the processed speech on each unseen noise at 0dB SNR are given in Figs. 12 and 13, respectively.

As can be seen from the results in Figs. 12 and 13, the overall performance of the processed

Table 3: Objective scores of different methods on unseen noise

	PESQ				**STOI**			
Mthods	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB
Noisy	1.38	1.51	1.66	1.83	0.66	0.72	0.78	0.84
Iter-KF	1.64	1.84	2.04	2.26	0.68	0.75	0.81	0.85
FNN-MAG	1.73	1.92	2.13	2.32	0.71	0.78	0.83	0.87
FNN-IRM	1.81	2.05	2.29	2.51	0.75	0.81	**0.86**	**0.90**
FSEGAN	1.74	1.95	2.16	2.35	0.69	0.76	0.82	0.85
FNN-KF	1.73	2.01	2.21	2.38	0.71	0.77	0.82	0.85
FNN-KFBE	1.74	2.03	2.25	2.43	0.72	0.79	0.84	0.87
Hybrid	**1.96**	**2.16**	**2.36**	**2.52**	**0.77**	**0.83**	**0.86**	0.89

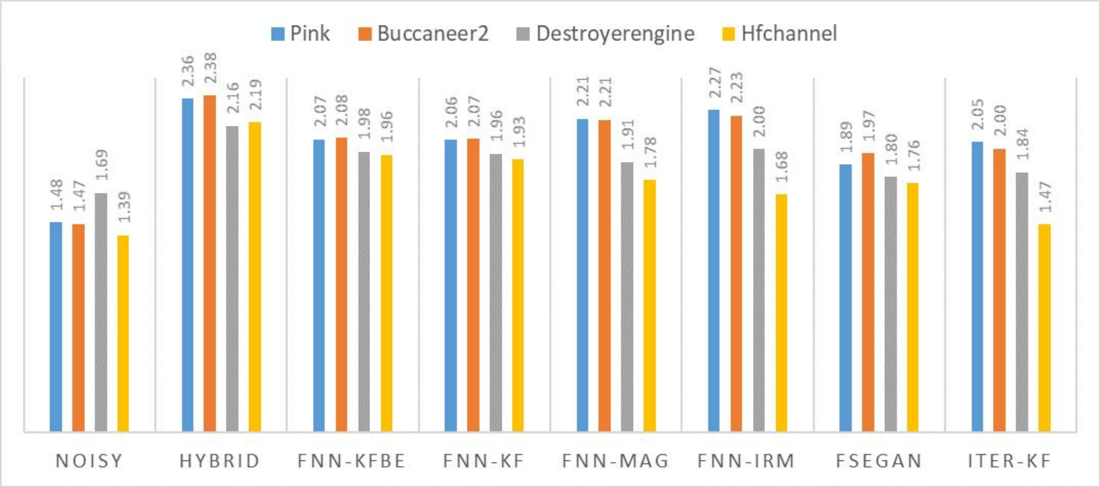


Figure 12: PESQ scores on different noises at 0dB SNR.

speech on pink and buccaneer noises is better than that obtained on destroyerengine and hfchannel noises for all the methods. This is because the former two unseen noises share more similarities with some of the training noises and exhibit a less complex structure when compared to the latter two noises, so that the FNN can output a more accurate prediction. This finding indicates that the performance of DNN-based methods indeed varies with different noises. According to the PESQ scores in Fig. 12, the proposed hybrid system produces enhanced speech with better quality for all test noises. Further, for the STOI scores, the proposed hybrid system still achieves the highest scores on all noises.

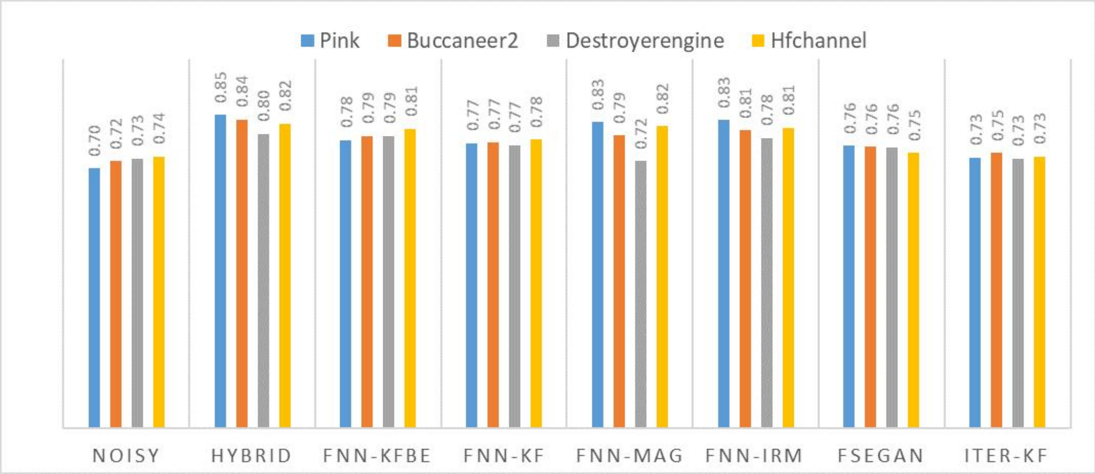


Figure 13: STOI scores on different noises at 0dB SNR.

3.5.4 Waveforms and Spectrograms of the Enhanced Speeches

In order to better understand the characteristics of the enhanced speech signals resulting from the methods under evaluation, illustrative waveforms and spectrograms are plotted and compared. The noisy speech is obtained by mixing a selected clean speech utterance with hfchannel noise at 3dB SNR.

Fig. 14 shows the residual noises and the distortions existing in the enhanced speech in the time domain. The processed speech from FSEGAN or FNN-MAG contains a large amount of residual noises, which is caused by the difficulty in learning the mapping from the noisy magnitude spectrogram to the clean one. Iter-KF and FNN-KF perform well in removing the additive noise, but they both bring distortion to the original speech. For example, the speech component after 0.3s is suppressed by Iter-KF while the magnitudes of the processed speech of FNN-KF is strongly attenuated. Note that the improvement of FNN-KFBE is not obvious by comparing the waveforms of FNN-KF and FNN-KFBE, more information will be presented in the comparison of the spectrograms. FNN-IRM and the proposed hybrid system achieve a better performance than the other methods, as they can remove more noise without bringing significant distortions. Finally, for this particular experiment with unseen noise, the hybrid system is slightly better than FNN-IRM, as the residual noise is lower in the unvoiced part near the middle of the utterance.

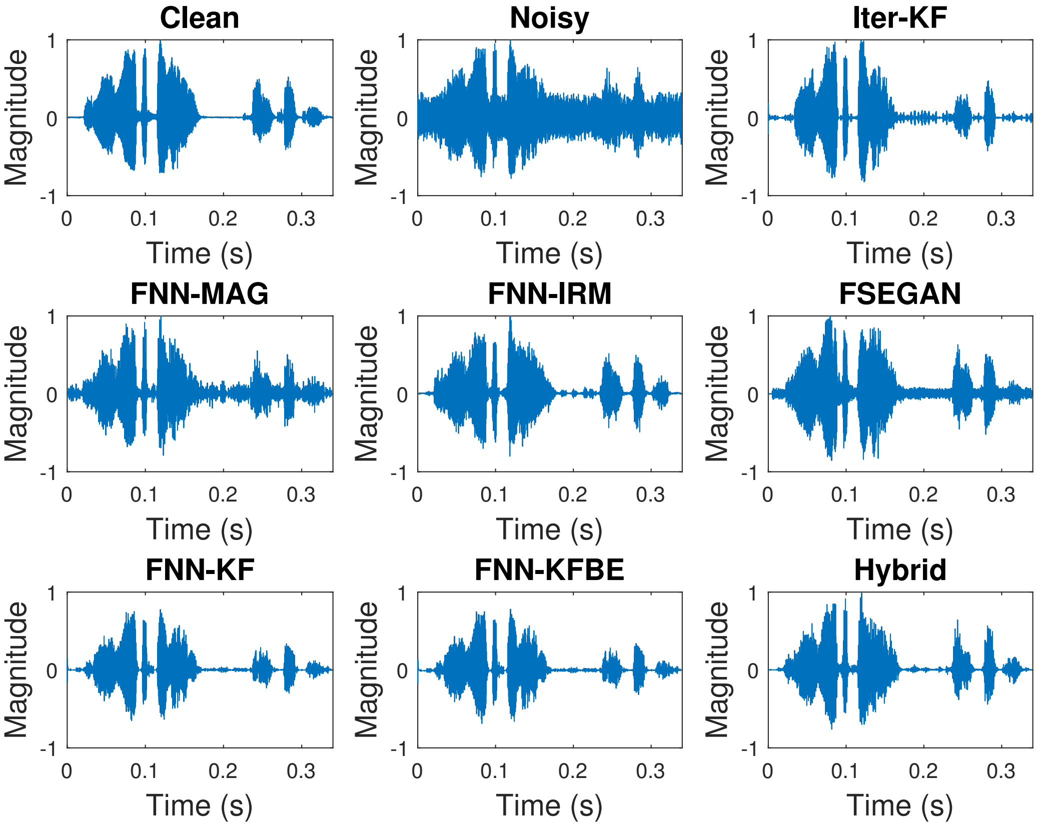


Figure 14: Time domain waveforms of the clean, noisy and enhanced speech signals for different methods.

Fig. 15 demonstrates the effects of the residual noises and the distortions in the harmonic structures of the enhanced speech in the T-F domain. For Iter-KF, we can see the musical noise structure in the spectrogram in the region between 2kHz and 3kHz. The spectrogram of FSEGAN also exhibits some undesirable structures, which likely cause the degradation of performance. We make further comparison among the four DNN-related methods. While the harmonic structures of the voiced parts with FNN-MAG are well preserved up to about 3kHz, a significant amount of residual noise is present during the unvoiced parts. The processed speech with FNN-IRM is affected by high-level residual broadband noise, which the method cannot adequately remove. While introducing less noise during the unvoiced parts, FNN-KF tends to suppress the HF components of the voiced parts of speech, leading to a decrease of speech quality. This problem could be alleviated by FNN-KFBE, whose spectrogram clearly shows that the HF component of the enhanced speech has been partly restored. Finally, the spectrogram of the enhanced speech with the proposed hybrid system seems to provide the best quality, i.e., clearer harmonic structures of the voiced part, and the fewer residual noises during unvoiced parts.

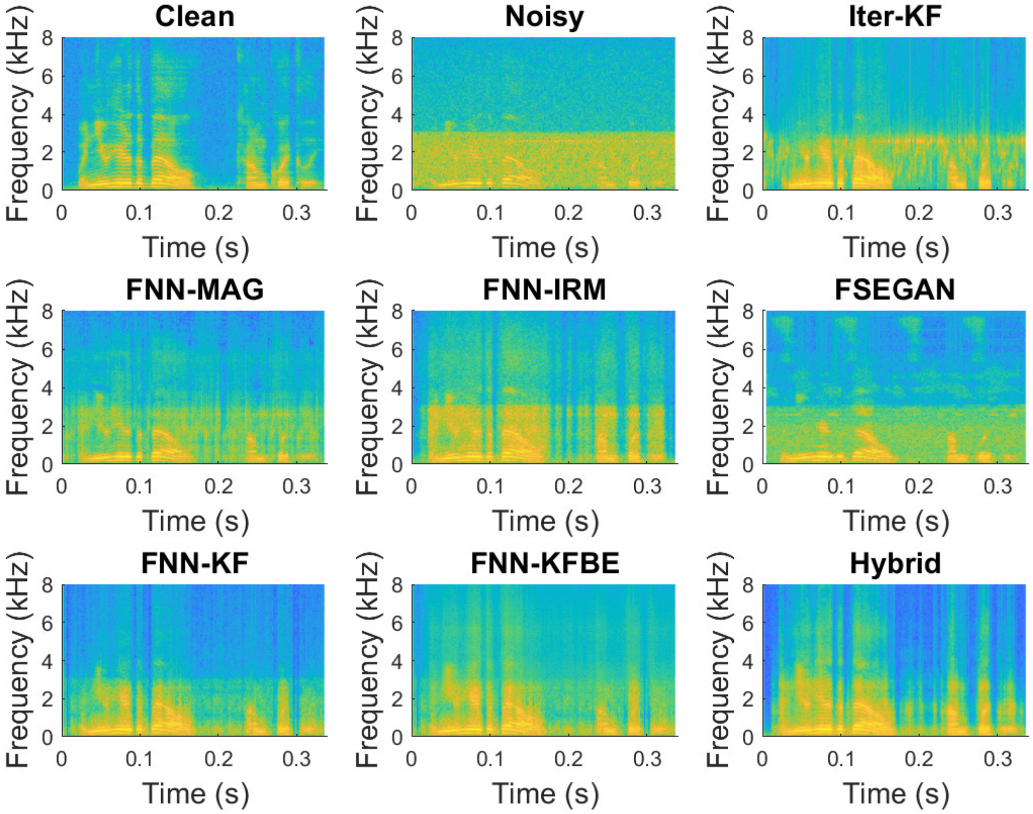


Figure 15: Spectrograms of the clean, noisy and enhanced speech signals for different methods.

3.6 Conclusion

In this chapter, we have first presented a DNN-augmented basic KF for time domain SE, in which an FNN is trained to estimate LPCs of clean speech model. With the help of the powerful learning capability of the DNN, our proposed system offers more accurate parameters for Kalman filtering, and thus achieves better performance than traditional iterative Kalman filtering. We further have applied a DNN-based bandwidth extension scheme for the HF component restoration of Kalman filtered speech, in order to compensate the speech degradation in the filtering process. Next, we have proposed a two-level hybrid SE system that combines DNN-based speech reconstruction and Kalman filtering. The first level aims to reconstruct the speech with estimated magnitude and noisy phase in T-F domain, and the second level further removes the residual noise in the reconstructed speech with Kalman filtering in time domain. Performance evaluation shows that the enhanced speech from the hybrid system has the best quality among all tested methods. Moreover, the proposed three systems embrace a better generalization capacity under unmatched noisy environment compared to existing DNN-based SE methods.

Chapter 4

DNN-Augmented Advanced Kalman Filter for Speech Enhancement

4.1 Introduction

In previous chapter, the DNN-augmented basic KF has been introduced and achieved good denoising performance. However, one limitation can not be ignored is that the HF component of the enhanced speech has been auttenuated. This is because the adoption of basic KF in the denoising process. To this end, several advanced versions of KFs have been proposed. The first category is the subband Kalman filtering technique [123, 124], which divides the noisy speech into several contiguous frequency bands, and performs Kalman filtering separately in each band. Subband Kalman filtering makes use of the fact that the noise does not affect the speech signal uniformly over the whole spectrum, and removes the noise with respect to the different frequencies to achieve better performance. The second category is the colored-noise KF [128, 129], which models both clean speech and noise as AR processes. This kind of method does not require the additive noise variance estimation in basic KF, and works better in color noise environments. Another category of the advanced KF is the perceptual KF [130, 131], which incorporates an additional post-filter to further remove the residual noise, by scaling the estimation error of the KF below the masking threshold. The threshold is calculated based on the perceptual theories of human ear system.

In this chapter, we extend our DNN-based parameter estimation technique to advanced KFs. In particular, we apply the DNN-based parameter estimation technique to the subband KF and colored-noise KF, respectively, in order to provide more accurate parameters for Kalman filtering and achieve better performances. This chapter is organized as follows. Section 4.2 presents a subband KF with DNN-estimated parameters, which contains three main parts: subband analysis and synthesis, Kalman filtering and DNN-based parameter estimation. In Section 4.3, we propose SE system based on a DNN-augmented colored-noise KF with spectral subtraction as post-processing. Performance of the proposed DNN-augmented advanced KFs are evaluated in Section 4.4 in terms of objective performance measures. Conclusions are drawn in Section 4.5.

4.2 Subband Kalman Filtering with DNN-Estimated Parameters

Subband analysis is widely adopted in speech processing, such as speech coding. It has also been applied to SE to separately reduce the background noise in different subbands [23, 123, 124, 132]. Experimental results of [123] and [23] have shown that the subband methods yield better performances compared to their respective full-band counterparts [19, 115].

In light of the successes of the previous subband techniques, we propose a novel DNN-augmented subband Kalman filtering system for SE, where the noisy speech is divided into subband speeches using discrete wavelet transform (DWT). For each noisy subband speech, the DNN is employed for the estimation of AR parameters and the KF is then applied to obtain the enhanced subband speech. The inverse DWT (iDWT) is finally used to obtain the enhanced full-band speech. Compared with the DNN-augmented basic KF, the DNN-augmented subband KF performs denoising at each subband, and is thus able to not only suppress the background noise but also reduce the speech distortion in the enhanced speech, especially at higher frequencies.

The overall block diagram of our DNN-assisted subband Kalman filtering system is depicted in Fig.16. It contains four parts: subband analysis, DNN-based LSFs estimation, Kalman filtering and subband synthesis. The details of each part are introduced in the following subsections.

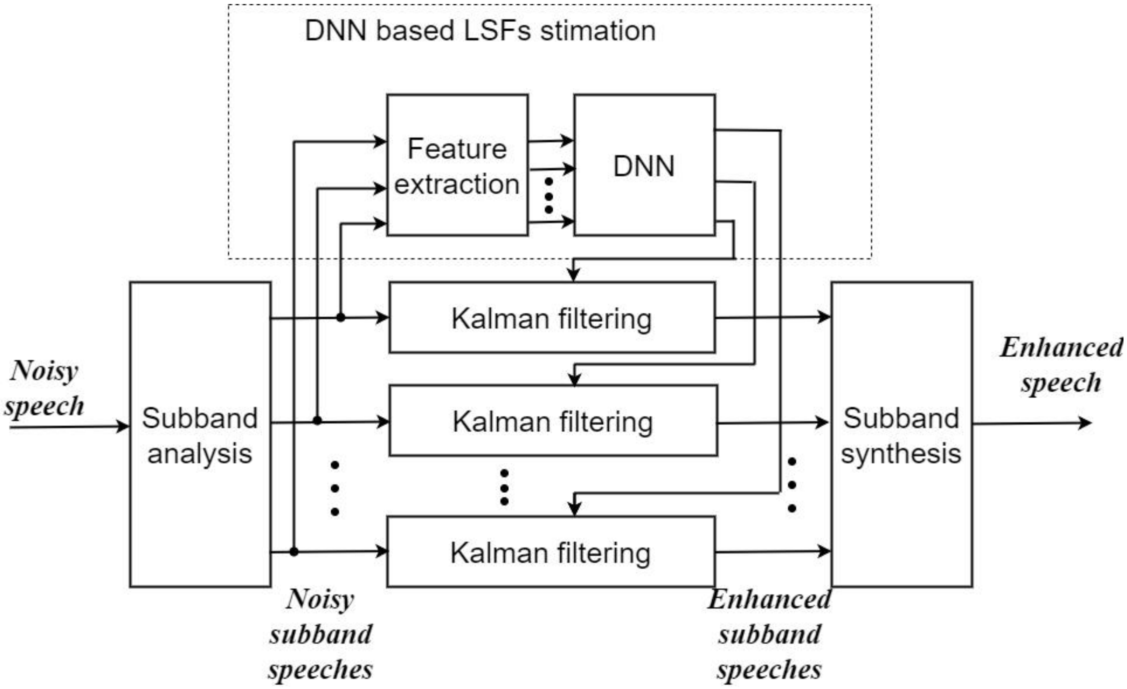


Figure 16: Block diagram of the subband Kalman filtering system.

4.2.1 Subband Analysis and Synthesis

Since the KF is viewed as a time domain estimator, we adopt the DWT to directly decompose the time domain noisy speech. This way can avoid the STFT operation, which brings moderate distortion to the time domain signal due to the necessary segmentation and windowing processes [132].

The DWT and iDWT are performed by a set of well-defined low-pass/high-pass filters together with a down/up-sampling process, which are regarded as distortionless analysis/synthesis for a time domain signal. Taking a 2-level DWT for an example, in the first level, DWT decomposes a full-band signal x into two subband signals with respect to the low and HF information components. In the next level, the decomposition operation is further applied to the LF subband signal, while the HF subband remains untouched. That is, with a J-level of DWT, we will obtain $J+1$ subband signals, as denoted by

$$x_b = DWT^J\{x\}, \quad b = 1, 2, \cdots, J+1, \tag{19}$$

where x_b is the b-th subband signal produced by DWT with b denoting the subband index.

Similarly, for subband synthesis, the iDWT is adopted to reconstruct a full-band signal $\hat{x}$ from the subband signals, which is given by

$$\bar{x} = iDWT^J\{\mathbf{x}\}, \tag{20}$$

where $\mathbf{x}$ denotes the set of all subband signals $\{x_b\}_{b=1}^{J+1}$. The reconstructed signal $\hat{x}$ is identical to the original signal x in the perfect reconstruction case.

4.2.2 Kalman filtering for Each Subband

While the noisy speech $y(n)$ is decomposed into subband speeches $\{y_b(n)\}_{b=1}^{J+1}$ in the subband analysis, KF is then applied to each noisy subband speech for denoising. The process of Kalman filtering for each noisy subband speech $y_b\,(n)$ is identical to the process introduced in Section 3.2.1. Given the corrupted subband speech $y_b(n)$, the KF recursively calculates an unbiased and linear MMSE estimate of the state vector $\mathbf{s}_b(n)$ as illustrated in Eq. (11). Three parameters should be determined beforehand, that is, the additive noise variance σ_w^2, the driving noise variance σ_v^2, and the transition matrix $\mathbf{F}$ with the LPCs of the clean subband speech. The additive noise variance is estimated and updated during the unvoiced frames and the driving noise variance is given by Eq. (15) The estimation of the LPCs of $s_b(n)$ is introduced in the following subsection.

4.2.3 DNN-Based LSFs Estimation

To begin with, the LPCs are converted to the LSFs in DNN-based estimation. The DNN-based LSFs estimation is divided into off-line and on-line stages. In the off-line stage, a DNN is trained to learn the mapping between the acoustic features of noisy subband speeches and the LSFs of the clean counterparts. In the on-line stage, given the features of a noisy subband speech, the well-trained DNN predicts the LSFs of the clean subband speech. It should be mentioned that instead of training several DNNs for different subbands separately, we employ a single DNN for all the subband speeches to better exploit the relationships within different subbands as well as to reduce the computational and structural complexity.

The input feature set is the same as introduced in Section 3.2.2. The structure of the DNN is a five-layer FNN, which consists of one input layer, three hidden layers with 1024 units in each layer, and one output layer. The activation function used in the hidden layer is the ReLU, while a linear function is used in the output layer.

Back propagation is used to find the optimal weights and biases of the DNN to minimize the cost function, which is defined as the mean square error (MSE) between the reference LSFs and the estimated ones for all subbands,

$$E_r = \frac{1}{J+1} \sum_{b=1}^{J+1} \left\{ \frac{1}{M_b} \sum_{m=1}^{M_b} \left\{ \frac{1}{p} \sum_{i=1}^{p} \left[\hat{\omega}_{b,i}(m) - \omega_{b,i}(m) \right]^2 \right\} \right\} \tag{21}$$

where M_b denotes the total number of frames for the b-th noisy subband speech, $\omega_{b,i}(m)$ and $\hat{\omega}_{b,i}(m)$ are the reference and the estimated LSFs for frame m, respectively, where $i \in \{1, ..., p\}$ is the order index of the clean speech AR model.

In summary, the proposed DNN-assisted subband Kalman filtering system includes an off-line training stage and an on-line enhancement stage. The former trains a DNN with subband noisy and clean speech pairs, while the latter is described in detail below.

- Decompose the full-band noisy speech $y(n)$ into the subband versions $\{y_b(n)\}$ with DWT.
- Extract the features of each noisy subband speech and employ the trained DNN to obtain the estimated LSFs, which are converted to the LPCs to form the transition matrix.
- Estimate the additive noise variance σ_w^2 during unvoiced frames and compute the driving noise variance σ_v^2 using Eq. (15).

- Perform Kalman filtering with Eq. (11) for each noisy subband speech $y_b(n)$ to obtain the enhanced counterpart $\hat{s}_b(n)$.

- Synthesize the enhanced subband speeches $\{\hat{s}_b(n)\}$ to reconstruct the final enhanced speech $\hat{s}(n)$ with iDWT.

4.3 Multi-Objective DNN-Augmented Colored-Noise Kalman Filter

Although in Section 3.2, the performance of the basic KF method for SE has been improved notably by using the FNN for parameter estimation, several limitations have been identified. Firstly, the additional VAD procedure needed for the estimation of the additive noise variance increases the computational and structural complexity of the system. In addition, accurately detecting the unvoiced frames remains a difficult task, and the detection errors lead to inaccurate variance estimation of the additive noise, which brings further distortion to the enhanced speech.

To counter the difficulties posed by the VAD procedure and improve the accuracy of the variance estimation, we propose a SE system that implements DNN-based parameter estimation to the colored-noise KF. The overall block diagram of our new system is depicted in Fig.17, which is composed of two stages: the training stage and the enhancement stage. In the training stage, the input feature set to the DNN consists of the combination of the noisy speech LSFs along with four acoustic features from [82]. The output targets are the LSFs of both the clean speech and the noise. Then, a multi-objective DNN is trained to learn the mapping from the noisy input feature set to the targets. In the enhancement stage, given a noisy speech signal, we obtain first the input feature set, and then process it by the trained DNN to predict the clean speech LSFs and noise LSFs. The estimated LPCs are then obtained from the LSFs, and applied to both variance estimation and Kalman filtering. Subsequently, the noisy speech is enhanced by the colored-noise KF. This operation is followed by a post subtraction to further remove the residual noise in the filtered speech. The key components and steps involved in the proposed system are described in further details below.

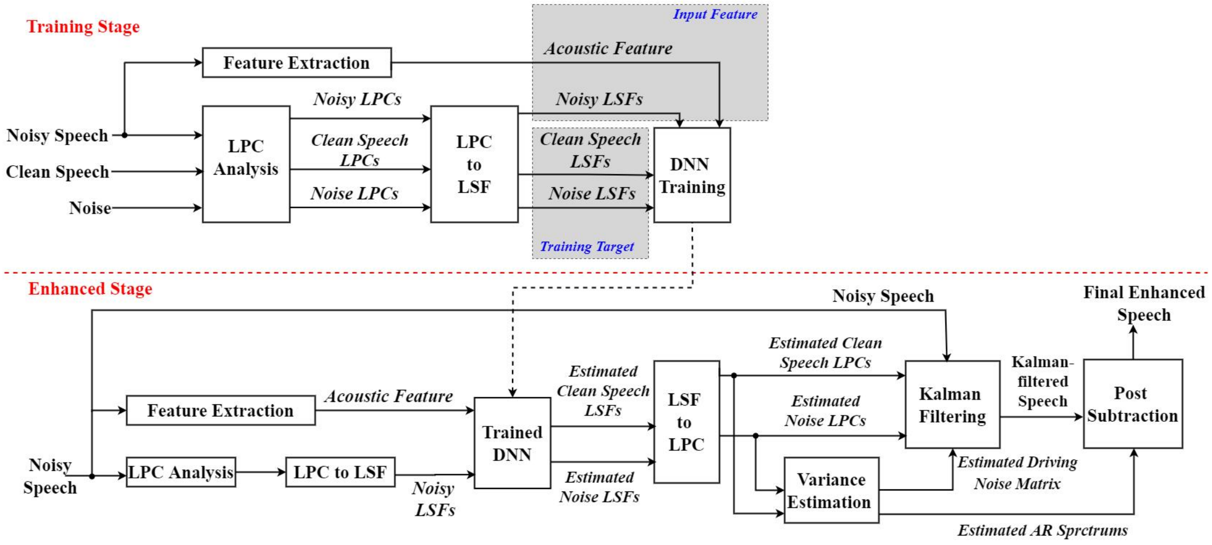


Figure 17: Block diagram of proposed SE system using DNN-augmented colored noise KF.

4.3.1 Colored-Noise Kalman filter

As mentioned before, in a conventional KF the clean speech is modelled as an AR process, while the additive noise is assumed to be white, which is not suitable for the complex noises encountered in real-world environment. To overcome this limitation, we herein adopt the colored-noise KF. In this method, the additive noise $w(n)$ in (1) is now modelled as an AR process, expressed as,

$$w\left(n\right)=\sum_{i=1}^{q}a_{w,i}w\left(n-i\right)+z\left(n\right) \tag{22}$$

where $a_{w,i}$ are the LPCs of the colored noise, q the order of the AR model, and $z(n)$ the zero-mean white driving noise with variance σ_z^2.

The underlying AR signal model in the colored-noise KF can be conveniently incorporated into the following state-space matrix form,

$$\begin{aligned}\mathbf{x}(n)&=\mathbf{F}\mathbf{x}(n-1)+\mathbf{G}\mathbf{u}(n)\\ y(n)&=\mathbf{H}^T\mathbf{x}(n)\end{aligned} \tag{23}$$

where $\mathbf{x}(n) = \Big[\mathbf{s}(n), \mathbf{w}(n)\Big]^T$ is the $p+q$ dimensional concatenated state vector constituted by the clean speech vector $\mathbf{s}\left(n\right) = \left[s\left(n-p+1\right), \ldots, s\left(n-1\right), s\left(n\right)\right]$ together with the noise vector $\mathbf{w}\left(n\right) = \left[w\left(n-q+1\right), \ldots, w\left(n-1\right), w\left(n\right)\right]$, and $\mathbf{u}(n) = \Big[v(n), z(n)\Big]^T$ is the concatenated driving noise vector. Moreover, the augmented matrices $\mathbf{G}$, $\mathbf{H}$, and the overall transition matrix $\mathbf{F}$ are given as follows:

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_w \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \mathbf{G}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_w \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_w \end{bmatrix} \tag{24}$$

with

$$\mathbf{F}_w = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ a_{w,q} & a_{w,q-1} & \cdots & a_{w,2} & a_{w,1} \end{bmatrix} \tag{25}$$

and $\mathbf{H}_w = \mathbf{G}_w = \left[0, \cdots, 0, 1\right]^T \in \mathbb{R}^q$.

Given a noisy observation $y(n)$, the estimate of the state vector $\mathbf{x}(n)$ can be obtained by the following Kalman filtering recursive equations:

$$\begin{cases} e\left(n\right) = y\left(n\right) - \mathbf{H}^T\hat{\mathbf{x}}\left(n|n-1\right) \\ \mathbf{K}\left(n\right) = \mathbf{P}\left(n|n-1\right)\mathbf{H}\left(\mathbf{H}^T\mathbf{P}\left(n|n-1\right)\mathbf{H}\right)^{-1} \\ \hat{\mathbf{x}}\left(n|n\right) = \hat{\mathbf{x}}\left(n|n-1\right) + \mathbf{K}\left(n\right)e\left(n\right) \\ \mathbf{P}\left(n|n\right) = \left(\mathbf{I} - \mathbf{K}\left(n\right)\mathbf{H}^T\right)\mathbf{P}\left(n|n-1\right) \\ \hat{\mathbf{x}}\left(n+1|n\right) = \mathbf{F}\hat{\mathbf{x}}\left(n|n\right) \\ \mathbf{P}\left(n+1|n\right) = \mathbf{F}\mathbf{P}\left(n|n\right)\mathbf{F}^T + \mathbf{G}\mathbf{Q}_u\mathbf{G}^T \end{cases} \tag{26}$$

where $e\left(n\right)$ is the innovation, $\mathbf{K}\left(n\right)$ the Kalman gain matrix, $\hat{\mathbf{x}}\left(n|n\right)$ the filtered estimate of state vector $\mathbf{x}\left(n\right)$, $\hat{\mathbf{x}}\left(n|n-1\right)$ *a priori* estimate of the state vector $\mathbf{x}\left(n\right)$. $\mathbf{P}\left(n|n\right)$ is the filtered state error covariance matrix, and $\mathbf{P}\left(n|n-1\right)$ the predicted state error correlation matrix. $\mathbf{Q}_u$ is the covariance matrix of the driving noise vector $\mathbf{u}(\mathbf{n})$, which is given by

$$\mathbf{Q}_u = E[\mathbf{u}(n)\mathbf{u}(n)^T] = \begin{bmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_z^2 \end{bmatrix}. \tag{27}$$

The denoised speech is the output of the colored-noise KF, i.e.,

$$\hat{s}\,(n) = [\mathbf{G}_s^T, \mathbf{0}^T]\hat{\mathbf{x}}\,(n|n) \tag{28}$$

Note that two parameterized matrices that appear in the process equations (26) should be estimated from the noisy speech to carry out Kalman filtering, namely, the overall transition matrix $\mathbf{F}$ and the covariance matrix $\mathbf{Q}_u$ of the concatenated driving noise vector $\mathbf{u}(n)$. The first depends on the clean speech and noise LPCs, which can be converted to the LSFs and predicted through a DNN, while the second one is obtained by solving an optimization problem. The details of this parameter estimation are provided in the following subsections.

4.3.2 DNN-Based LSFs Estimation

Recently, in [133], we have demonstrated that FNN offers a convenient means for LSFs estimation in speech processing applications. Here, we propose to employ two different networks, i.e., FNN and LSTM, to predict both the clean speech LSFs and noise LSFs. The specific configuration of FNN is described in Section 3.2.2. The LSTM network is obtained by stacking by one input layer for high-level feature extraction, two LSTM layers with 512 units in each layer for temporal modelling, one feed-forward layer with 512 units and one output layer for LSFs estimation.

The input feature set is the same as presented in Section 3.2.2. For the training targets, we adopt a multi-objective learning architecture to estimate both the clean speech LSFs and noise LSFs. Compared to a standard DNN, the output layer in the proposed architecture is divided into two parts: one for the clean speech LSFs and the other for the noise LSFs. The advantages of multi-objective learning are twofold. On one hand, it has lower computational complexity compared to training two separate DNNs (i.e., one for clean speech and one for noise). On the other hand, estimating the two sets of LSFs simultaneously can help better exploit the relationship between the clean speech and noise.

In the training stage, back propagation is used to adjust the weights and biases so as to minimize the cost function, which is defined as the mean square error (MSE) between the reference LSFs and the estimated ones for each training utterance. Note that the cost function is composed of two parts: one for the clean speech LSFs and the other for the noise LSFs, as given by,

$$\begin{aligned} \mathrm{MSE}_{\mathrm{LSF}} =& \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{1}{p}\sum_{i=1}^{p}\left[\hat{\omega}_{s,i}\left(m\right)-\omega_{s,i}\left(m\right)\right]^{2}\right. \\ & \left.+\frac{1}{q}\sum_{j=1}^{q}\left[\hat{\omega}_{w,j}\left(m\right)-\omega_{w,j}\left(m\right)\right]^{2}\right\} \end{aligned} \tag{29}$$

where m is the frame index of the input noisy speech and M the total number of the frames. The quantities $\omega_{s,i}(m)$ and $\hat{\omega}_{s,i}(m)$ are the reference clean speech LSFs and the estimated ones at frame m, where $i \in \{1, ..., p\}$ is the order index of the clean speech AR model. Similarly, $\omega_{w,j}$ are the reference noise LSFs and $\hat{\omega}_{w,j}$ the estimated ones at frame m, where $i \in \{1, ..., q\}$ is the order index of the noise AR model.

In the enhancement stage, the clean speech LSFs and noise LSFs are first obtained by the well-trained DNN, and then converted to their respective LPCs. The estimated LPCs are used along with the estimated variances in the KF equations (26) in order to estimate the desired speech signal.

4.3.3 Variance Estimation

The covariance matrix $\mathbf{Q}_u$ in (27) is another key parameter that needs to be estimated prior to the application of the Kalman filtering equations. Proceeding as in [134], we now formulate an optimization problem to estimate σ_v^2 and σ_z^2. Our goal is to minimize the difference between the noisy spectrum and the sum of the estimated clean speech spectrum and noise spectrum.

From equations (1), (8) and (22), the spectrum of the AR-modelled noisy speech can be expressed as:

$$\begin{aligned}\hat{P}_y(k) &= \hat{P}_s(k) + \hat{P}_w(k) \\ &= \frac{\sigma_v^2}{|A_s(k)|^2} + \frac{\sigma_z^2}{|A_w(k)|^2}\end{aligned} \tag{30}$$

with

$$\begin{aligned}A_s(k) &= 1 - \sum_{i=1}^{p} a_{s,i} e^{-j2\pi ik/K} \\ A_w(k) &= 1 - \sum_{i=1}^{q} a_{w,i} e^{-j2\pi ik/K}\end{aligned} \tag{31}$$

where K is the frame length. Note that the clean speech LPCs $a_{s,i}$ and the noise LPCs $a_{w,i}$ can be obtained from the LSFs at the output of the trained DNN.

The AR spectrum of the observed noisy speech $P_y(k)$ can be written as,

$$P_y(k) = \frac{\sigma_y^2}{|A_y(k)|^2} \tag{32}$$

with

$$A_y(k) = 1 - \sum_{i=1}^{p} a_{y,i} e^{-j2\pi ik/K} \tag{33}$$

$$\sigma_y^2 = E\left[y(n)^2\right] - \mathbf{r}_y^T\, \mathbf{a}_y. \tag{34}$$

We can obtain the variance estimates by minimizing the difference between the AR spectrum of the modelled noisy speech $\hat{P}_y(k)$ and that of the observed one $P_y(k)$, that is,

$$\sigma_v^{*2}, \sigma_z^{*2} = \text{arg}\min_{\sigma_v^2, \sigma_z^2} d\left(\hat{P}_y(k), P_y(k)\right) \tag{35}$$

where the difference is measured in the log-spectral domain as given by,

$$\begin{aligned} d\left(\hat{P}_y(k), P_y(k)\right) &= \frac{1}{K}\sum_{k=1}^{K}\left|\ln\hat{P}_y(k) - \ln P_y(k)\right|^2 \\ &\approx \frac{1}{K}\sum_{k=1}^{K}\left|\frac{\sigma_v^2/\left|A_s(k)\right|^2 + \sigma_z^2/\left|A_w(k)\right|^2 - P_y(k)}{P_y(k)}\right|^2. \end{aligned} \tag{36}$$

To obtain the approximate equation in (36), we have used equation (30) and the approximation of $\ln(x+1) \approx x$. Then by applying partial differentiation to the difference $d\left(\hat{P}_y(k), P_y(k)\right)$ with respect to σ_v^2 and σ_z^2, we obtain the following linear system of equations:

$$\begin{bmatrix} E_{ss} & E_{sw} \\ E_{sw} & E_{ww} \end{bmatrix}\begin{bmatrix} \sigma_v^2 \\ \sigma_z^2 \end{bmatrix} = \begin{bmatrix} E_{ys} \\ E_{yw} \end{bmatrix} \tag{37}$$

with

$$\begin{aligned} E_{ss} &= \left\|\frac{1}{P_y^2(k)\left|A_s(k)\right|^4}\right\|, E_{ww} = \left\|\frac{1}{P_y^2(k)\left|A_w(k)\right|^4}\right\| \\ E_{sw} &= \left\|\frac{1}{P_y^2(k)\left|A_s(k)\right|^2\left|A_w(k)\right|^2}\right\| \\ E_{ys} &= \left\|\frac{1}{P_y(k)\left|A_s(k)\right|^2}\right\|, E_{yw} = \left\|\frac{1}{P_y(k)\left|A_w(k)\right|^2}\right\| \end{aligned} \tag{38}$$

The norms involved in equation (38) are defined as $\|f(k)\| \triangleq \sum_{k=1}^{K}|f(k)|$.

When the AR spectrum of the observed noisy speech $P_y(k)$ is calculated and $A_s(k)$ and $A_w(k)$ are obtained with the estimated LPCs from the trained DNN, we can finally obtain the optimal variances σ_v^2 and σ_z^2 using equation (37).

4.3.4 Post Subtraction

To further remove the residual noise in the Kalman-filtered speech, a post subtraction algorithm is applied right after Kalman filtering. We adopt multiband spectral subtraction because of its good performance in reducing speech distortion [23]. The main idea of this method is described as follows.

The fast Fourier transform (FFT) is first applied to the windowed Kalman-filtered speech to obtain the magnitude spectrum. Next, the noise spectrum is estimated and updated during the unvoiced frames. The detection of unvoiced frames is accomplished by comparing the total power of the estimated clean speech, say $\hat{P}_s^2$ and that of the estimated noise, $\hat{P}_w^2$, which can easily be obtained from the estimated spectra in Section 4.3.3. Specifically, a frame is labelled as a voiced frame if $\hat{P}_s^2 > \hat{P}_w^2$, and as an unvoiced frame otherwise.

Then, the magnitude spectra of the filtered speech and noise are divided into L subbands. In each subband, the Kalman-filtered magnitude spectrum is enhanced by subtracting a noise power spectrum term,

$$|\hat{C}_l\,(k)\,|^2 = |\hat{S}_l\,(k)\,|^2 - \alpha\,\delta_l\,|\hat{D}_l\,(k)\,|^2 \tag{39}$$

where $|\hat{C}_l\,(k)\,|^2$ denotes the modified subband speech power spectrum, $|\hat{S}_l\,(k)\,|^2$ the Kalman-filtered speech power spectrum and $|\hat{D}_l\,(k)\,|^2$ the estimated noise power spectrum (obtained and updated during unvoiced frames), with k being the discrete frequency and l the subband index. Moreover, α is the oversubtraction factor and δ_l the additional subtraction factor that can be individually set for each subband to customize the noise removal process.

The factors α and δ_l are used to control the noise subtraction level within each band. The value of α is defined as a function of the SegSNR (in dB), i.e.,

$$\alpha = \begin{cases} 4.75 & , \quad \text{SNR} < -5 \\ 4 - \dfrac{3}{20}\text{SNR} & , \quad -5 \le \text{SNR} \le 20 \\ 1 & , \quad \text{SNR} > 20 \end{cases} \tag{40}$$

with

$$\text{SNR} = 10\log_{10}\left(\frac{|\hat{S}_l\,(k)\,|^2}{|\hat{D}_l\,(k)\,|^2}\right) \tag{41}$$

The value of δ_l is determined as,

$$\delta_l = \begin{cases} 1 & , \quad f_l < 1\text{kHz} \\ 2.5 & , \quad 1\text{kHz} \leq f_l \leq \dfrac{F_s}{2} - 2\text{kHz} \\ 1.5 & , \quad f_l > \dfrac{F_s}{2} - 2\text{kHz} \end{cases} \tag{42}$$

where f_l is the upper frequency of the l-th band and F_s the sampling frequency. The above values of the factors α and β are taken from [23] where they have been determined empirically based large experiments.

Finally, we synthesize the modified subband spectrum from the modified magnitude (39) and the phase of the Kalman-filtered speech. The final enhanced speech is obtained by computing the inverse FFT of the modified subband spectrums.

4.4 Experimental Results

4.4.1 Experimental Setup

The clean speech is selected from the IEEE sentence database [99], while the noise is from the NOISEX-92 database [108]. For the preparation of the pairs of noisy speech and clean speech, we adopt the same setting as described in Section 3.5.1. PESQ and STOI are adopted to evaluate the enhancement performance, with a higher score means a better speech quality or intelligibility, respectively.

4.4.2 Reference Methods

To evaluate the proposed SE system, we adopt several existing methods for performance comparison. Our reference methods include both KF based algorithms and DNN-based approaches. The following provides a brief conceptual summary of each one of the reference methods.

- P-IKF (Perceptual iterative KF) [131]: This algorithm calculates a perceptual mask according to human hearing system and applies it to the Kalman-filtered speech in order to further remove the residual noises.

- S-IKF (Subband iterative KF) [124]: In this method, the noisy speech is first divided into subband signals. Iterative Kalman filtering is then applied separately for each subband noisy speech. The final enhanced speech is obtained by synthesising the subband enhanced speech signals.

- FNN-MAG [96]: An FNN is employed to directly explore the mapping from the noisy speech magnitude spectrum to the clean one. The enhanced speech is synthesised with the estimated clean magnitude and noisy phase.

- FNN-WF [37]: An FNN is trained for the estimation of AR parameters of the clean speech. Then, a Wiener filter is estimated by calculating the ratio of the estimated clean speech power spectrum to that of the noisy speech. The enhanced speech is then obtained by applying the estimated Wiener filter to the noisy speech.

- FNN-KF [133]: An FNN is used to predict the LPCs needed for conventional Kalman filtering. The DNN learns the mapping from the acoustic features of the noisy speech to the LSFs of the clean speech. The estimated LSFs are then converted to the desired LPCs.

Besides these benchmarks, we consider the proposed subband Kalman filtering and three versions of the proposed DNN-augmented colored-noise KF method, namely,

- FNN-SKF: Subband Kalman filtering system using FNN for LSFs estimation

- FNN-CKF: Colored-noise Kalman filtering system using FNN for LSFs estimation while without post subtraction.

- FNN-CKFS: Colored-noise Kalman filtering system using FNN for LSFs estimation and with post subtraction.

- LSTM-CKFS: Colored-noise Kalman filtering system using LSTM for LSFs estimation and with post subtraction.

In order to make fair comparisons, we use the same configuration for the FNN in the related methods as mentioned in Section 3.5.1. For FNN-MAG, a Hamming window is selected to divide each utterance into 20 ms time frames with a 10 ms frame shift (50% overlap). A 320-point DFT is then computed for each frame. For the other reference methods and the proposed system, a rectangular window is used to divide the audio signals into 20 ms frames with no overlap.

For the conventional KF, we set $\mathbf{s}(0|0) = \mathbf{0}$, $\mathbf{P}(0|0) = \mathbf{I}$, and the AR model order of the clean speech as $p = 12$. For the colored-noise KF, we set $\mathbf{x}(0|0) = \mathbf{0}$, $\mathbf{P}(0|0) = \mathbf{I}$, and the orders of AR models for clean speech and additive noise as $p = q = 12$. For the post subtraction in FNN-CKFS and LSTM-CKFS, the spectrum is evenly divided into 4 bands. For the subband analysis in FNN-SKF, we use the Symlets 13 for DWT.

4.4.3 Level of Subband Analysis

First, we decompose the noisy speech at different levels to find the optimal subband analysis level J in the FNN-SKF system. Three levels ($J = 1, 2, 3$) are tested under seen noise. Table 4 shows the objective results under the assumption that the clean speech is accessible to obtain ideal AR parameters for Kalman filtering. In this case, the three subband KFs outperform the full-band processing, which indicates that denoising in each subband indeed removes the additive noise better and introduces less speech distortion. In addition, decomposing the speech with a deeper level contributes to a better performance when the ideal parameters are available.

Table 5 shows the objective results where the DNN-estimated AR parameters are employed for Kalman filtering. We find that adopting 1-level DWT for subband analysis, namely decomposing the noisy speech into two subband speeches, leads to the best result. As the subband analysis level gets deeper, the number of the input subband signals is increased, which requires a more complex structure to perfectly learn the relationships between more input features and the targets. As such, the 1-level led to better enhancement result. Another possible reason is that if we decompose at a deeper level, more KFs are required. Since the parameters cannot be ideally estimated for each KF, the estimation error leads to a degradation of the enhanced subband speeches. Thus, the synthesized full-band speech suffers more performance decrease for high-level subband analysis cases. As a result, we choose 1-level decomposition for the FNN-SKF system.

Table 4: Objective results using ideal AR parameters for subband Kalman filtering under seen noise

		-3dB	0dB	3dB	6dB
PESQ	Noisy	1.41	1.52	1.68	1.86
	Full-band processing	2.37	2.54	2.70	2.86
	1-level analysis	2.38	2.55	2.72	2.90
	2-level analysis	2.39	2.56	2.74	2.91
	3-level analysis	**2.40**	**2.57**	**2.75**	**2.93**
STOI	Noisy	0.66	0.72	0.78	0.83
	Full-band processing	0.84	0.87	0.89	0.90
	1-level analysis	0.86	0.88	0.90	0.92
	2-level analysis	0.87	0.89	0.91	0.92
	3-level analysis	**0.88**	**0.90**	**0.92**	**0.93**

Table 5: Objective results using estimated AR parameters for subband Kalman filtering under seen noise

		-3dB	0dB	3dB	6dB
PESQ	Noisy	1.41	1.52	1.68	1.86
	Full-band processing	1.70	1.93	2.13	2.30
	1-level analysis	**1.92**	**2.16**	**2.36**	**2.57**
	2-level analysis	1.81	2.05	2.27	2.50
	3-level analysis	1.68	1.88	2.12	2.36
STOI	Noisy	0.66	0.72	0.78	0.83
	Full-band processing	0.71	0.77	0.81	0.85
	1-level analysis	**0.72**	**0.78**	**0.84**	**0.87**
	2-level analysis	0.71	0.77	0.83	**0.87**
	3-level analysis	0.69	0.75	0.82	0.86

4.4.4 Evaluation of LPCs Estimation Accuracy

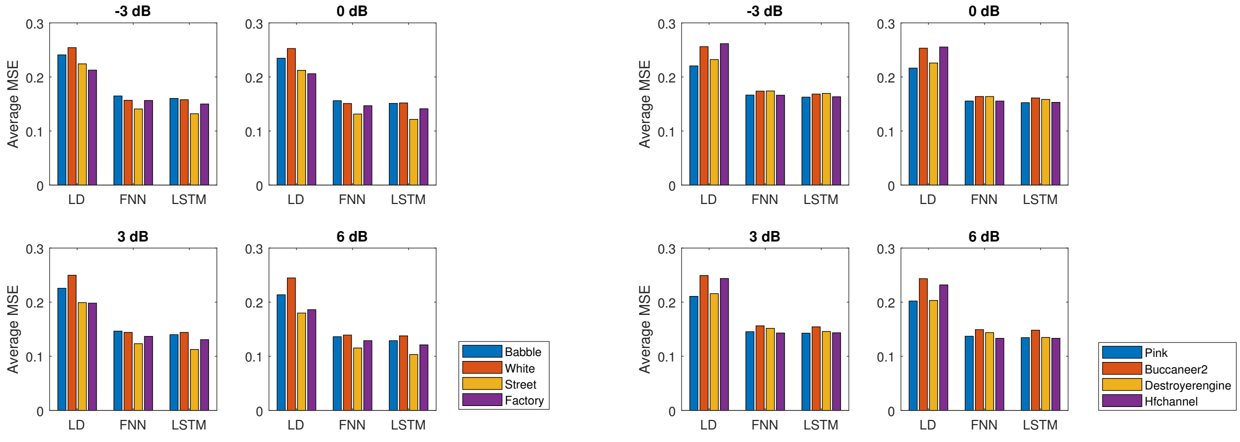
In this subsection, the LPCs estimation error is evaluated to verify the learning capability of the proposed multi-objective training in DNN-augmented colored-noise Kalman filtering system. We first define the LPCs estimation error of the speech as the mean square error (MSE) between the estimated LPCs and the ideal LPCs calculated from the clean speech for each utterance as given below,

$$\mathrm{MSE}_{\mathrm{LPC}} = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{1}{p}\sum_{i=1}^{p}\left[\hat{a}_{s,i}\left(m\right)-a_{s,i}\left(m\right)\right]^{2}\right\} \tag{43}$$

where M denotes the number of the speech frames in the utterance, $a_{s,i}(m)$ the ideal LPCs of the clean speech and $\hat{a}_{s,i}(m)$ the estimated ones. The estimated LPCs are obtained by three methods for comparison. The first one applies the Levinson-Durbin (LD) algorithm to obtain the LPCs of the noisy speech directly [135]. The second and third ones adopt the proposed DNN-based LSFs estimation algorithm, where FNN and LSTM are used to estimate the LSFs, which are then converted to LPCs. Similarly, we compute the LPCs estimation error of the additive noise for each noise type by using (43), where the estimated and ideal LPCs of the speech are replaced by those of the additive noise, and the order of the speech model, p is replaced by that of the noise, q.

Fig. 18 shows the LPCs estimation error comparison for the speech. The average MSE is computed over all the testing utterances for both seen and unseen noise. In general, the FNN and LSTM based approaches give a slightly smaller error than the LD method does for the eight types of noises and different SNRs. In addition, the error from LSTM is smaller than that from FNN in most cases. Another important finding is that the errors from the DNN methods decrease with an increase of the SNR, which means that DNN achieves a better performance at higher SNR. The LPCs estimation performance also varies for different noise types. In particular, the best estimation accuracy is achieved for street noise, and the worst for white noise. Interestingly, we note that the estimation error of FNN and LSTM based algorithm under unseen noise does not increase considerably compared with that under seen noise, which indicates that using DNN in LPCs estimation offers robustness and has a good generalization capability.

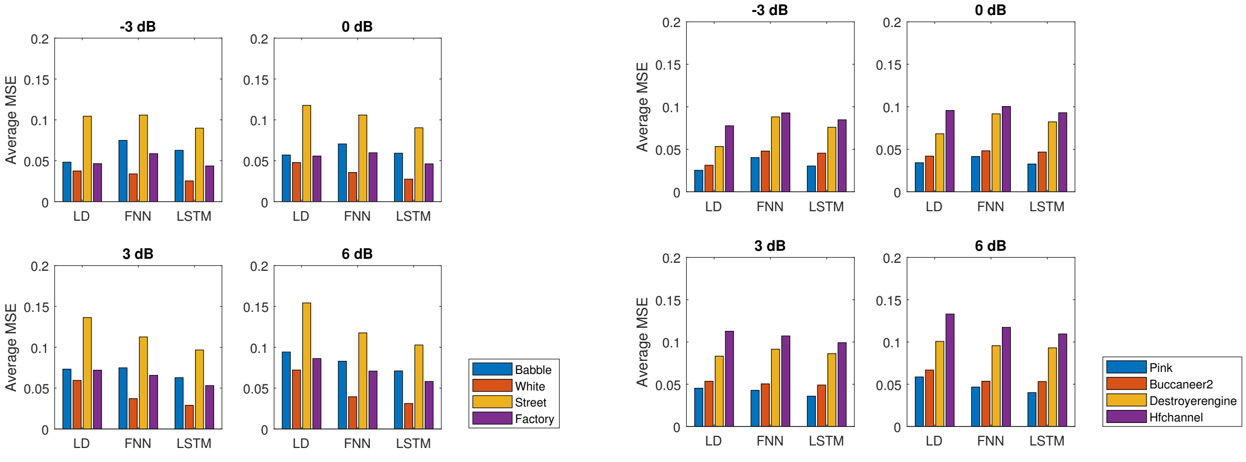
The LPCs estimation error comparison for the additive noise is shown in Fig. 19, where we notice important differences with the case of clean speech. Firstly, as SNR increases, the strong speech component more strongly affects the noise LPCs estimation, and hence the LPCs estimation



(a) Seen noise (b) Unseen noise

Figure 18: LPC estimation error comparison for speech among Levinson-Durbin (LD), FNN and LSTM methods.

error of additive noise gets larger. Secondly, compared with speech, noise exhibits less structure and correlation, and the mapping from the noisy speech feature to the noise LSFs is thus more difficult to learn. Therefore, we can find that the DNN estimation result is not always better than the traditional LD method, especially at low SNR.



(a) Seen noise (b) Unseen noise

Figure 19: LPC estimation error comparison for additive noise among LD, FNN and LSTM methods.

4.4.5 Speech Enhancement Performance under Seen Noise

Here, we compare the different SE methods under seen noise. Table 6 gives the average objective scores of different SE methods on seen noise. We first note that the performances of the unsupervised Kalman filtering algorithms are worse than those of the DNN-based methods. The P-IKF, which incorporates a perceptual mask to further suppress the residual noise, is the best among the three unsupervised Kalman filtering algorithms. However, P-IKF still can not achieve as good performance as FNN-KF, not to mention our FNN-SKF, FNN-CKF, FNN-CKFS and LSTM-CKFS. These results demonstrate the benefit from employing DNN in parameter estimation. The DNN can predict more accurate LPCs from the noisy speech, thus improving the performance of Kalman filtering algorithms.

Moreover, FNN-KF has lower objective scores compared with the proposed methods. This is because FNN-KF requires a VAD procedure to detect the unvoiced frame for estimating and updating the additive noise variance σ_w^2. However, VAD in noisy condition is a difficult task, which causes variance estimation error and introduces extra distortion to the enhanced speech. Compared with FNN-KF, although FNN-SKF also requires VAD for the additive noise variance estimation, the subband processing enables FNN-SKF to adaptively remove the additive noise in each band, thus contributes to obtaining a higher objective score. For the colored-noise Kalman filtering system, an AR model is adopted to represent the background noise. As such, the Kalman filtering equations in (26) no longer involve σ_w^2, and we can therefore overcome the speech distortion problem due to the inaccurate estimation of σ_w^2. The performance can be further improved by employing post subtraction to remove the residual noise due to the inaccurate parameters of the noise AR model. Indeed, FNN-CKFS achieves a better performance than FNN-CKF, which approaches closely that of FNN-MAG. Comparing FNN-SKF with FNN-CKFS, we notice that the FNN-SKF has higher PESQ scores, while the FNN-CKFS has higher STOI scores, which indicates that the enhanced speech from FNN-CKFS has a better intelligibility. Finally, although FNN-MAG has the best performance among all tested FNN-based approaches, by employing LSTM for LSFs estimation in the proposed system, LSTM-CKFS can achieve the best PESQ scores, which demonstrates the LSTM's advantage in modelling long temporal dependencies.

Table 6: Objective scores of different SE methods under seen noise

Method	**PESQ**				**STOI**			
	-3 dB	0 dB	3 dB	6 dB	-3 dB	0 dB	3 dB	6 dB
Noisy	1.41	1.52	1.68	1.86	0.66	0.72	0.78	0.83
P-IKF	1.57	1.83	2.08	2.31	0.68	0.75	0.81	0.85
S-IKF	1.56	1.81	2.04	2.29	0.67	0.75	0.81	0.84
FNN-MAG	1.89	2.13	2.34	2.55	**0.75**	**0.82**	**0.86**	**0.88**
FNN-WF	1.65	1.83	2.15	2.36	0.71	0.78	0.82	0.86
FNN-KF	1.70	1.93	2.13	2.30	0.71	0.77	0.81	0.85
FNN-SKF	1.92	2.16	2.36	2.57	0.72	0.78	0.84	0.87
FNN-CKF	1.73	2.01	2.26	2.49	0.72	0.78	0.84	0.87
FNN-CKFS	1.88	2.12	2.32	2.51	0.73	0.79	0.85	**0.88**
LSTM-CKFS	**1.93**	**2.16**	**2.38**	**2.58**	0.74	0.80	0.85	**0.88**

4.4.6 Speech Enhancement Performance under Unseen Noise

Table 7 gives the average objective scores of the different SE methods in the case of unseen noise. In this case, the performances of the unsupervised Kalman filtering algorithms are still worse than those of the FNN-based methods. Comparing FNN-KF with our proposed system, we can find again that FNN-SKF, FNN-CKF, FNN-CKFS and LSTM-CKFS outperform FNN-KF because of the adoption of colored-noise KF. However, the STOI scores of FNN-CKF are slightly lower than those of FNN-KF at low SNR. This degradation is possibly caused by the inaccuracy in estimating the noise LPCs, as shown in Fig.19 where the FNN estimation error is higher than the LD estimation error under low input SNR conditions.

In the case of unseen noise, we find that LSTM-CKFS achieves the best objective scores due to its advanced network structure. More interestingly, FNN-MAG no longer holds the best performance among FNN-based methods. In fact, the objective scores of FNN-MAG decease largely, indicating that mapping the noisy magnitude spectrum to the clean one is prone to errors when the noise is unmatched with those in the training stage. In contrast, FNN-WF, FNN-KF and our

Table 7: Objective scores of different SE methods under unseen noise

Method	**PESQ**				**STOI**			
	-3 dB	0 dB	3 dB	6 dB	-3 dB	0 dB	3 dB	6 dB
Noisy	1.37	1.51	1.65	1.82	0.65	0.72	0.78	0.83
P-IKF	1.67	1.88	2.09	2.32	0.69	0.76	0.81	0.85
S-IKF	1.66	1.87	2.08	2.31	0.68	0.75	0.81	0.84
FNN-MAG	1.73	1.92	2.13	2.32	0.70	0.76	0.82	0.87
FNN-WF	1.68	1.92	2.15	2.33	0.67	0.74	0.81	0.85
FNN-KF	1.73	1.95	2.21	2.38	0.71	0.77	0.82	0.85
FNN-SKF	1.87	2.10	2.33	2.55	0.71	0.78	0.83	0.88
FNN-CKF	1.76	2.02	2.26	2.48	0.70	0.76	0.82	0.86
FNN-CKFS	1.89	2.11	2.32	2.50	0.71	0.78	0.83	0.87
LSTM-CKFS	**1.91**	**2.15**	**2.36**	**2.55**	**0.73**	**0.79**	**0.84**	**0.88**

proposed advanced Kalman filtering system suffer less performance degradation. Indeed, the denoising process in these methods is accomplished by Wiener and Kalman filtering. Therefore, as the DNN can provide more accurate parameters, their performances would not fluctuate as much whether on seen noise or unseen noise. Based on these results, and considering the robustness of the DNN-based LPCs estimation, we can conclude that our FNN-SKF, FNN-CKF, FNN-CKFS and LSTM-CKFS, have a better generalization capability than FNN-MAG.

Finally, we make comparison in terms of each objective metric. Although the enhanced speech from LSTM-CKFS has the best speech quality according to the PESQ scores, the improvement of speech intelligibility is not obvious as seen from the STOI scores. In fact, the LSTM-CKFS gives similar STOI scores to FNN-MAG. Actually, there is a trade-off between residual noise and speech distortion for SE algorithms, leading to decreased speech intelligibility. For our LSTM-CKFS, the enhanced speech achieves similar speech intelligibility as that of FNN-MAG but far better speech quality, indicating that LSTM-CKFS could preserve the information content of clean speech well, while significantly removing the additive noise.

4.4.7 Spectrograms of Enhanced Speeches

To better understand the characteristics of the enhanced speech, Fig. 20 shows the spectrograms of the enhanced speech signals from several selected methods, demonstrating the effects of the residual noises and the distortions in the harmonic structures in the T-F domain. The noisy speech is obtained by mixing a selected clean speech utterance with buccaneer noise at 3dB SNR. For the best unsupervised Kalman filtering in our experiment, i.e., P-IKF, we can find the musical noise structure in the spectrogram in the region between 4kHz and 8kHz. The spectrogram of FNN-MAG also exhibits some musical noise structures in the HF component as well as residual noise in the LF component. For FNN-WF, the HF components look better than the previous two spectrograms, but still have undesired structures, which are likely caused by the difficulty of Wiener filter in removing non-stationary noise. Finally, for the five Kalman filtering related methods, it is observed that FNN-KF, FNN-CKFS and LSTM-CKFS can remove the background noise quite well. However, the HF components of FNN-KF still suffer from various degradations. While this situation is improved in the cases of FNN-SKF, FNN-CKFS and LSTM-CKFS. Such as FNN-SKF, its denoising process is separately conducted according to the noise level in each subband, thus it preserves the HF component better. However, the amount of the residual noise is larger compared to FNN-KF and FNN-CKFS. Finally, the LSTM-CKFS can preserve the harmonic structures best among all the tested methods, thus achieving the best objective scores.

4.5 Conclusion

In this chapter, we have extended the idea of DNN-based parameter estimation to the advanced KFs. First, we have presented a subband KF with DNN-estimated parameters for time domain SE, in which the noisy speech is divided into several subband, and Kalman filtering is then applied to each subband signals separately. Next, based on the knowledge that both clean speech and noise can be modelled as AR processes, we have proposed a DNN-augmented colored-noise KF system with post subtraction to better counter the non-stationary noise.

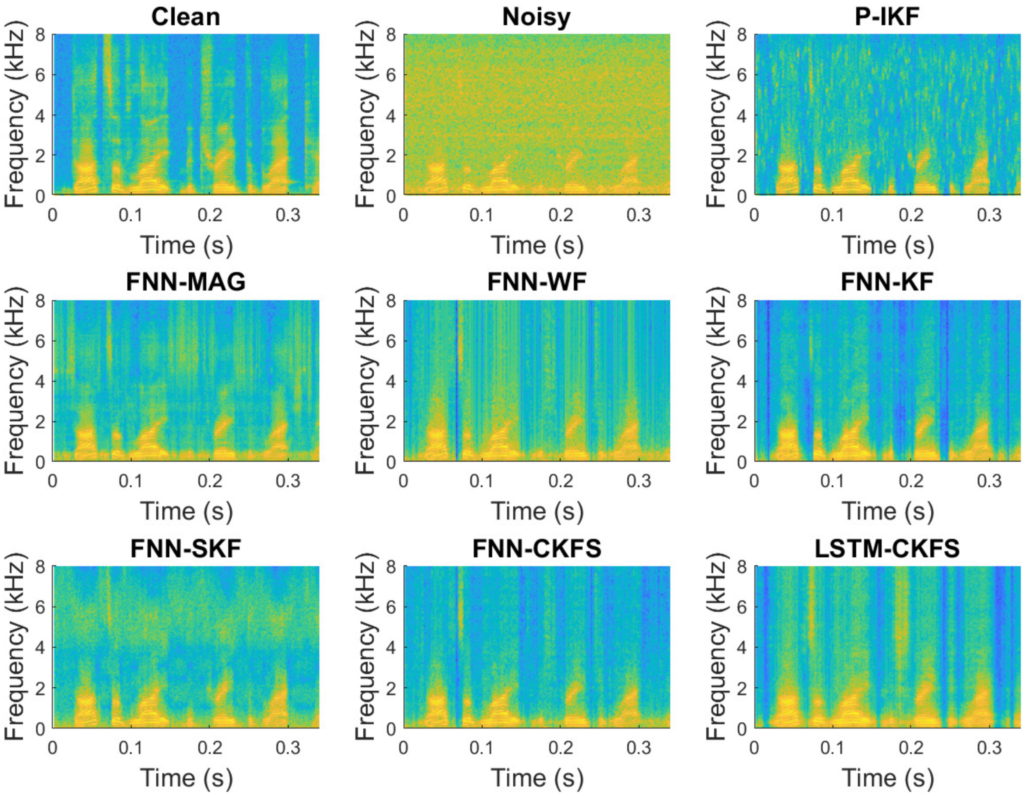


Figure 20: Spectrograms of the clean, noisy and enhanced speech signals for different methods

Performance comparison is conducted among tradition KF based methods, current DNN-based methods, and our proposed DNN-augmented basic KF, subband KF and colored-noise KF systems. The experimental results reveal that applying DNN to advanced KF provides accurate parameters in filtering and thus improves the enhancement performance. Moreover, the DNN-augmented subband KF shows a considerable advantage, since subband analysis allows the filter to individually remove the background noise according to the noise level in each subband. The DNN-augmented colored-noise KF achieves the best performance because the noise signal is also considered as an AR process, which avoids the VAD procedure in basic KF. Besides, employing spectral subtraction as post processing helps further remove the residual noise.

Experiments on adopting different networks for parameter estimation are also conducted, which reveals the LSTM's superiority in modeling temporal dependencies. Finally, similar to the DNN-augmented basic KF, the DNN-augmented advanced KFs also show better generalization capability as compared to other tested DNN-based methods.

Chapter 5

Constrained Masking Techniques for Speech Enhancement

5.1 Introduction

In this chapter, we target the problem of SE by using a well-known category of methods, i.e., the masking techniques, which have received much attention and achieved a series of progresses [136]. Inspired by the masking effect of the human auditory mechanism, the goal of this kind of methods is to estimate a mask, which can be applied to the noisy speech to retain the speech-dominant regions and suppress the noise-dominant regions. An appropriate and accurate mask is of great importance to the enhancing performance. To this end, researchers have made large efforts from two aspects: designing an efficient mask and developing reliable mask estimation algorithms.

The IBM [137] is one of the pioneering masks investigated in the literature. Given a clean speech in a T-F representation, the mask value of a T-F unit is set to 1 if the local SNR is greater than a preset threshold, otherwise it is set to 0. This simplifies SE to a binary classification problem. However, SE using IBM has some limitations such as introducing the residual musical noise. As a result, the IRM [34, 36], which can be viewed as a smoothed form of IBM, is proposed. The IRM is obtained by computing the ratio between the energy of clean speech and that of noisy speech for each T-F unit. Denoising with IRM is actually assigning large ratios to the T-F units with higher local SNR and small ratios to those with lower local SNR. Another mask with similar concept is the spectral magnitude mask or ideal amplitude mask (IAM) [126], which computes the ratio of the magnitude of clean speech to that of noisy speech. Note that IRM and IAM are motivated by the frequency response of the Wiener filter, which achieves the optimal SNR gain for stationary signals. However, speech signals and many real-world noises are nonstationary. To overcome this problem, the optimal ratio mask (ORM) is proposed in [138], by considering the correlation between the desired speech and noise, leading to an improved SNR of the enhanced speech. The above mentioned masks only focus on enhancing the magnitude spectrogram. More recently, the phase information has been considered in masking techniques, such as the PSM [52] and cIRM [53], to better recover the complex speech spectrogram.

Probably, supervised learning algorithms, such as Gaussian mixture model [71] or support vector machine [139], are the most popular mask estimation methods in early works. In recent years, the deep learning based methods have made a great progress. An FNN is adopted in [36] to learn the mapping between noisy acoustic features and IRM. Other architectures can be found in [140, 141], where recurrent neural network and CNN are, respectively, employed as the estimation model. The deep structure and powerful learning capability enable DNN to better explore the non-linear relationship between noisy features and masks, leading to a better estimation result.

Although the DNN-estimated masks have achieved good performance in improving speech intelligibility, none of these works further investigates the trade-off between speech distortion and residual noise in the enhanced speech. Several traditional methods have been proposed to denoise with less speech distortion and remove residual noise as much as possible. In [142], the authors proposed a new weighting rule using masking properties, which calculates the weighting coefficients to keep the perceived noise to be equal to a pre-defined level. However, the speech distortion is not explicitly considered in their processing. In [59], a spectral constraint is applied into the STSA estimator, which adaptively suppresses the noise-dominant regions and reduces the speech distortion in the speech dominant regions.

In this chapter, we will develop a new type of mask, called constrained ratio mask, for SE. The rest of this chapter is organized as follows. Section 5.2 describes the proposed CRM and its application in SE. In Section 5.3, we extend the CRM to enhance the complex spectrogram of the speech signal. Performance assessment is presented in Section 5.4 and conclusions are drawn in Section 5.5.

5.2 Speech Enhancement with Constrained Ratio Mask

In this section, we propose a new type of mask, namely CRM, for SE, which is derived to minimize the speech distortion while suppressing the residual noise such that it falls below a threshold level. A DNN is then trained for CRM estimation, which learns a mapping from the noisy features to the CRM. Finally, the enhanced speech is obtained by applying the estimated CRM to the noisy speech. Compared with the previous mask based algorithms, which mainly focus on retaining the speech information, our CRM based system is the first one to consider both speech distortion and the residual noise level in the enhanced speech, by adaptively adjusting the value of the CRM for different T-F units according to their local SNRs.

5.2.1 Constrained Ratio Mask

As usual, we consider a noisy speech $y(n)$ as the addition of clean speech $s(n)$ and background noise $w(n)$, with n denoting the time index. The time domain noisy speech can be transformed into a spectro-temporal spectrogram using STFT, namely, $Y(k,l) = S(k,l) + W(k,l)$, where $Y(k,l)$, $S(k,l)$ and $W(k,l)$ denote the STFT spectrograms of the noisy speech, clean speech and noise, respectively, with k and l indicating the frequency bin and frame index. We denote the ratio mask as $M(k,l) \in [0,1]$, which will be applied into the magnitude of the noisy speech to get the enhanced magnitude, i.e., $|\hat{S}| = M \cdot |Y|$. For simplicity, we have omitted k and l without loss of generality. By directly using the noisy phase for speech reconstruction, we obtain the STFT of the enhanced speech,

$$\hat{S} = (M \cdot |Y|)\, e^{j\phi_y} = M \cdot Y \tag{44}$$

To derive a CRM, we first introduce the mathematical definitions of the speech and noise distortions as given in [142, 143]. The estimation error e of $\hat{S}$ can be decomposed into two items as follows,

$$\begin{aligned} e &= \hat{S} - S = M \cdot (S + W) - S \\ &= (M-1) \cdot S + M \cdot W \\ &= e_s + e_w \end{aligned} \tag{45}$$

The power spectrums of e_s and e_w can be written as,

$$\begin{aligned} d_s = \mathrm{E}\left[{e_s}^2\right] &= (M-1)^2 \cdot P_s \\ d_w = \mathrm{E}\left[{e_w}^2\right] &= M^2 \cdot P_w \end{aligned} \tag{46}$$

where P_s and P_w are the power spectrums of the clean speech and noise, respectively; d_s denotes the speech distortion and d_w the noise distortion. We regard d_s as the distortion to the original clean speech introduced by the enhancement algorithm, while d_w is the distortion caused by the residual noise. The above mentioned two distortion terms with respect to the value of mask M are plotted in Fig.21 for three different values of input SNR $\xi = P_s/P_w$.

Ideally, we prefer a large value of M to yield a small speech distortion d_s under the circumstance of the clean speech being much stronger than the noise (i.e., $\xi \gg 1$). Conversely, if the signal is weaker than the noise (i.e., $\xi \ll 1$), a small value of M is required so that the residual noise or d_w will be small. However, we found that when employing a DNN to estimate the IRM (tested under four different noises and input SNRs), the values of estimated IRM are $0.15\% \sim 5.67\%$ higher than those of reference IRM, which results in a larger d_w, indicating that the enhanced speech suffers more residual noise.

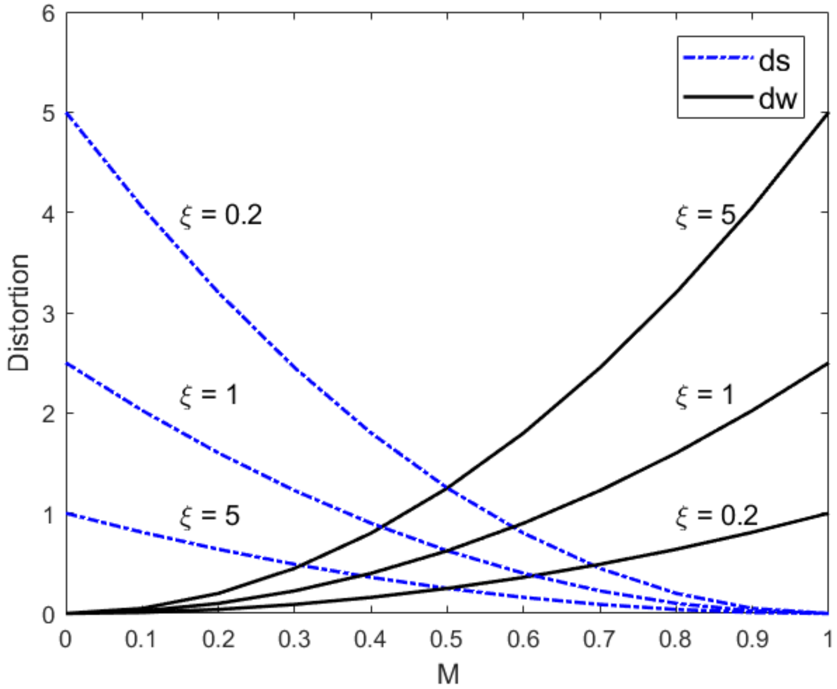


Figure 21: The relationship between M and distortions

To better remove the residual noise with no noticeable speech distortion, we need to derive a mask to minimize the speech distortion while constraining the noise distortion below a threshold. To this end, we establish the following constrained optimization problem,

$$\min_{M} d_s$$
$$\text{subject to}\, d_w \leq \delta \tag{47}$$

where δ is a preset threshold. It has been shown in [59] that the optimal M for (47) satisfies the following equation:

$$(M-1)\,P_s + \mu M P_w = 0 \tag{48}$$

where μ ($\mu \geq 0$) is the Lagrange multiplier (also named as the controlling factor in our paper).

From (48), the CRM can be expressed as

$$M = \frac{P_s}{P_s + \mu P_w} = \frac{\xi}{\xi + \mu} \tag{49}$$

It should be noted that due to the unknown SNR ξ, the controlling factor μ is to be determined. In other words, μ can be viewed as a function of the local SNR or ξ in dB. Using (49) into (46), the speech and noise distortions can be rewritten as,

$$\begin{aligned} d_s &= (M-1)^2 \cdot P_s = \left(\frac{\mu}{\xi + \mu}\right)^2 \cdot P_s \\ d_w &= M^2 \cdot P_w = \left(\frac{\xi}{\xi + \mu}\right)^2 \cdot P_w \end{aligned} \tag{50}$$

By adaptively adjusting the value of μ for each T-F unit, our CRM can balance the trade-off between the speech and noise distortions. For example, we would like to set a small value of μ for the speech dominant unit, in order to minimize the speech distortion and conserve the speech information; while for the noise dominant unit, a large value of μ is chosen to remove the noise as much as possible. As such, we propose the following empirical expression for μ:

$$\mu = \begin{cases} \mu_0 - \text{SNR}/s & , \quad S_l \leq \text{SNR} \leq S_u \\ \mu_{\min} & , \quad \text{SNR} > S_u \\ \mu_{\max} & , \quad \text{SNR} < S_l \end{cases} \tag{51}$$

where SNR $= 10\log_{10}\xi$, $\mu_{\max}$ and $\mu_{\min}$ are the maximum and minimum values of μ, respectively, S_l and S_u are the lower and upper bounds of the local SNR, respectively, and μ_0 and s are constants related to $\mu_{\max}$ and $\mu_{\min}$.

5.2.2 Proposed System with CRM

The proposed system is made of two stages: the off-line training stage and the on-line enhancement stage. In the training stage, a DNN is employed to learn the mapping between the noisy acoustic features and the reference CRM computed from speech databases. In the enhancement stage, given a new noisy speech, its features are extracted and sent to the well-trained DNN to obtain the CRM estimate, which is then applied to obtain the enhanced magnitude. Finally, the enhanced speech is reconstructed with the enhanced magnitude and noisy phase. The main steps involved in the proposed SE system are explained below.

DNN-Based CRM Estimation

In [82], the authors have investigated several acoustic features are investigated for supervised mask estimation and verified that using the feature set, which contains AMS, RASTA-PLP, MFCC, and GFCC, can obtain the best performance. Since these features lie in different ranges, normalization is required to scale the input features for achieving better results. Moreover, to make use of the temporal information of the speech, the features of two adjacent time frames are incorporated with the current frame to form a input feature set.

The architecture adopted is five-layer FNN that includes an input layer, an output layer and three hidden layers with 1024 units in each layer. We employ the ReLU as activation function in the hidden layers, and employ the linear function in the output layer.

To learn the weights and biases in the network, the famous back propagation is adopted to update the parameters in the training process. Ideally, the model parameters are trained to minimize the cost function J, which is defined as the mean square error between the reference and the estimated CRM.

$$J = \frac{1}{2L} \sum_{l=1}^{L} \sum_{k=1}^{K} \left(\hat{M}(k,l) - M(k,l) \right)^2 \tag{52}$$

where L denotes the total number of speech frames.

Waveform Reconstruction

In the enhancement stage, the estimated CRM is first output by the well-trained DNN. Afterwards, we apply the estimated CRM to the noisy magnitude spectrum to obtain the estimated magnitude. The enhanced speech spectrum is then reconstructed with the estimated magnitude $\hat{S}(k,l)$ and the noisy phase $\phi_y(k,l)$ as $\hat{S}(k,l) = |\hat{S}(k,l)|e^{j\phi_y(k,l)}$. Finally, the enhanced speech $\hat{s}(n)$ is obtained by performing the inverse STFT of $\hat{S}(k,l)$.

5.3 Extension of CRM for Complex Spectrogram Estimation

As explained in Section 1.3.2, phase processing has attracted more and more interests in SE. In [57], the authors proposed a complex spectrogram estimation system, which employs a multi-objective DNN to jointly estimate the mask: IRM, and the phase derivative: IFD. The estimated mask has two usages: estimating the clean magnitude and guiding the IFD in the phase reconstruction. To extend CRM for complex spectrogram estimation, we propose to updating the training target of the system in [57] as shown in Fig. 22. More specifically, we substitute the IRM with our proposed CRM, and investigate two phase derivatives: IFD and GD. Then, we evaluate the performance of the modified complex spectrogram estimation system under various noisy conditions. Since the procedure of DNN-based estimation and the magnitude processing are similar as presented in Section 5.2.2, the remaining of this section will focus on the explanation of phase reconstruction.

As Fig.23 (b) depicts, the phase spectrogram fluctuates rapidly along the time and frequency axes in contrast to the magnitude spectrogram, and the values of the phase are uniformly distributed due to phase wrapping [144]. The irregularities of the phase spectrogram cause difficulties in phase-aware enhancement by DNN-based approach. To overcome this barrier, a highly-structured new target derived from the phase is strongly required.

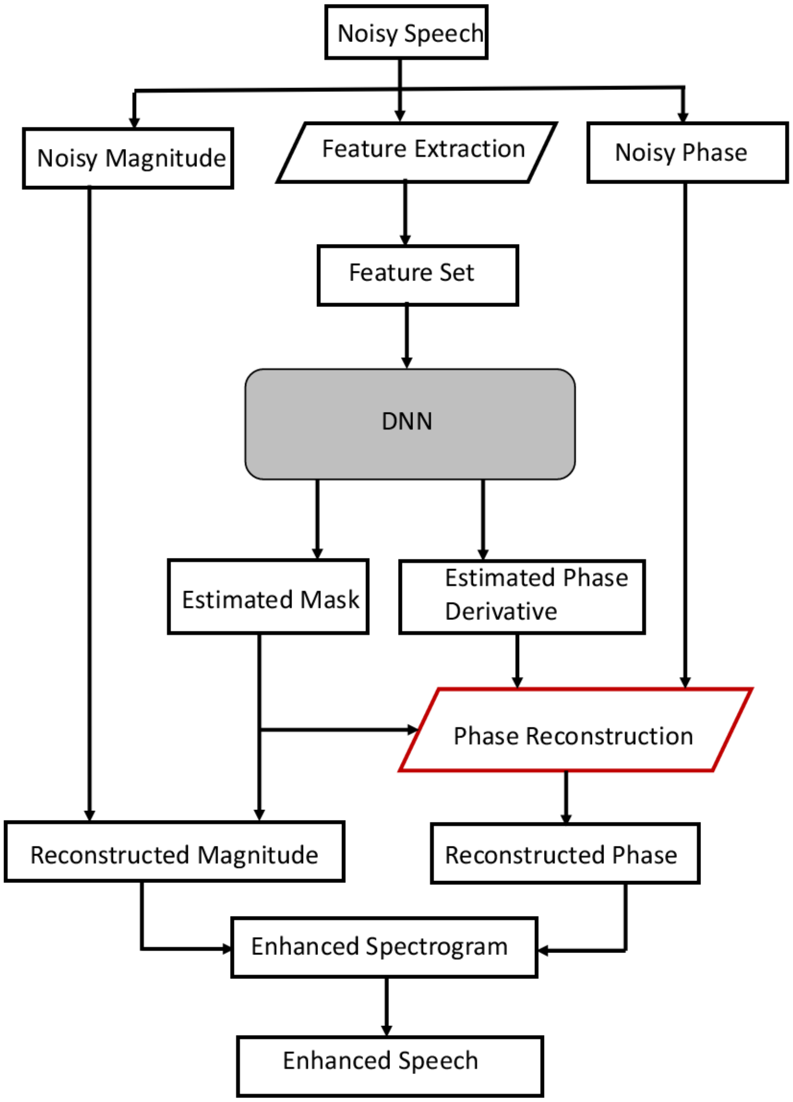
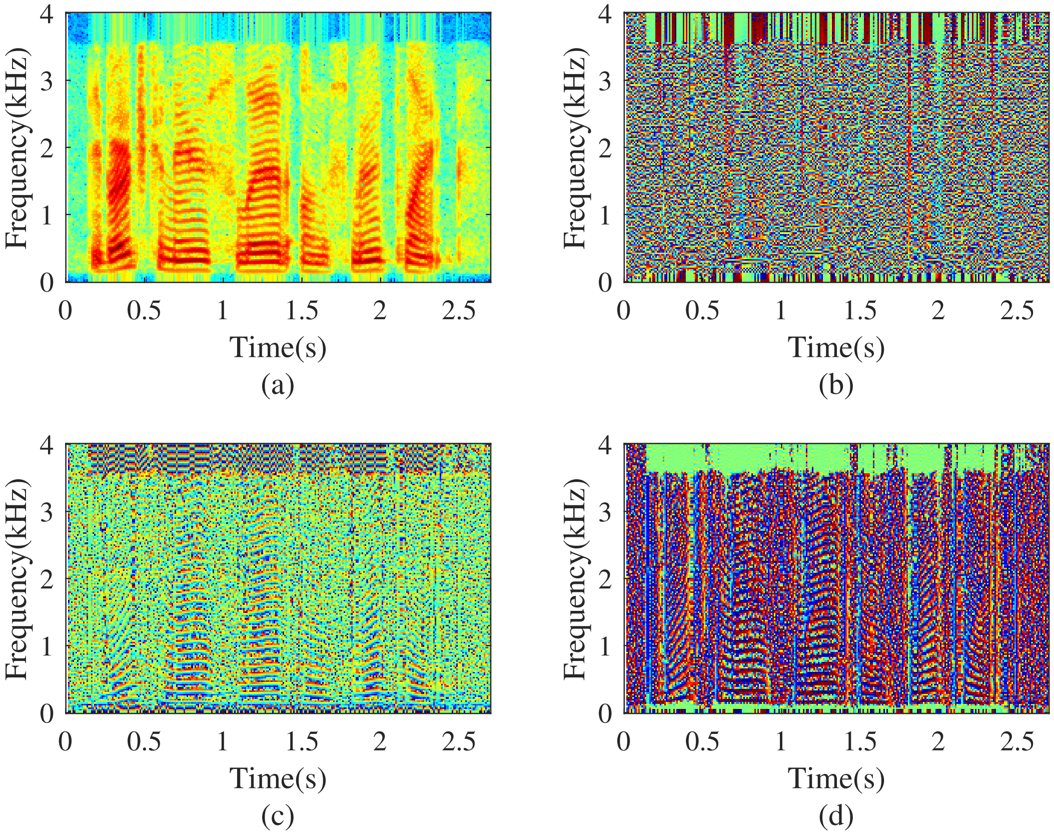


Figure 22: Block diagram of complex spectrogram estimation with mask and phase derivative

5.3.1 Phase Derivatives

Processing phase derivatives instead of phase itself are widely adopted in phase-aware SE. Among them, the instantaneous frequency (IF) [145] and group delay (GD) [146] are two of the most well-known phase derivatives.

IF is defined as the first time-derivative of the phase spectrum. For discrete signals, IF can be



(a)

(b)

(c)

(d)

Figure 23: Spectrogram plot of 3 seconds clean speech with sampling frequency 8 KHz, (a) Magnitude, (b) Phase, (c) IFD, (d) GD

approximated by the phase difference between two successive units:

$$\mathrm{IF}(k,l) = \mathrm{princ}\left\{\phi(k+1,l) - \phi(k,l)\right\} \tag{53}$$

where the function $\mathrm{princ}\{\cdot\}$ denotes the principal value operator, which projects the phase difference onto $[-\pi, \pi)$. Since the IF is limited to its principle value, the wrapping effects would occur along frequency axis. To alleviate the problem, the IFD is then adopted [145], which is given by,

$$\mathrm{IFD}(k,l) = \mathrm{IF}(k,l) - l \tag{54}$$

Basically, the IF values track the frequencies of pitch harmonic peaks and IFD magnitude denotes the accuracy of tracking. It is also demonstrated that IFD magnitude is inversely proportional to the spectral magnitude [145]. Therefore, the spectrogram of IFD captures the pitch and formant structure in a manner similar to the magnitude spectrogram. Similar findings are presented in [57], in which the authors reconstructed the phase with the estimated IFD for SE. They also proved that the IFD is able to be estimated with DNN, because IFD and the magnitude spectrogram have a similar pattern, which is shown in Fig. 23 (a) and (c).

GD is the negative derivation of the spectral phase with respect to frequency. In discrete case, GD is give by:

$$\mathrm{GD}(k,l) = -\left[\phi(k,l+1) - \phi(k,l)\right] \tag{55}$$

The GD can also be computed from the signal as in [146] using,

$$\mathrm{GD}(k,l) = \frac{S_R(k,l)X_R(k,l) + S_I(k,l)X_I(k,l)}{|S(k,l)|^2} \tag{56}$$

where the subscripts R and I denote the RI parts of the Fourier transform. $S(k,l)$ and $X(k,l)$ are the Fourier transform of $s(t)$ and $t(t)$, respectively. In [146], the authors demonstrated that the GD function behaves like a squared magnitude response at the resonance frequency. Similar structures can also be found by comparing (a) and (d) in Fig.23. Moreover, the high-resolution property discussed in [147] reveals that the GD function has a higher resolving power compared to the magnitude spectrum, i.e., the formants are resolved better in the group delay spectrum when compared to the magnitude or linear prediction spectrum. Based on this finding, we infer that GD can also be employed as a training target of DNN-based SE, just like the widely-adopted targets: magnitude or its variants.

5.3.2 Phase Reconstruction

Phase reconstruction is performed after obtaining the estimation phase derivatives by the well-trained DNN. Since the estimated phase derivatives are only the difference between the T-F units of the phase spectrogram along the time or frequency axis. Therefore, appropriate initial phase estimates in some T-F units is required to recover the phase spectrogram. Based on the initial estimates, the entire phase spectrogram can be reconstructed along the time and frequency axes with the estimated phase derivatives.

1) Initial phase estimation: It has been demonstrated that when the clean speech power is much larger than the noise power, the noisy phase is approximately equal to the clean phase. In other words, in high local SNR regions, using the noisy phases as initial estimate has a higher reliability. As suggested in [57], we also adopt the noisy phase spectrogram as the initial estimate of the clean phase spectrogram,

$$\hat{\phi}_{init}\left(k,l\right) = \phi_{y}\left(k,l\right), \forall k, \forall l. \tag{57}$$

Then, we use the local SNR of each T-F unit as index to determine the reliability of the initial estimate, where the local SNR is given by the estimated mask $\hat{M}(k,l)$.

2) Phase reconstruction with GD: At first, we must obtain the estimated GD by denormalizing the normalized GD. Then, the phase can be estimated by using the initial phase estimate together with the GD between the initial estimate and the target phase. For each T-F unit, we generate $(2N_s+1)$ frame-conditioned phase estimates, which is given by:

$$\hat{\phi}^{i}\left(k,l\right) = \begin{cases} \hat{\phi}_{init}\left(k,l{+}i\right){+}\sum\limits_{n=0}^{i-1}\hat{\text{GD}}\left(k,l{+}n\right), & i \neq 0 \\ \hat{\phi}_{init}\left(k,l{+}i\right), & i = 0 \end{cases}, \qquad \forall \quad -N_s \leq i \leq N_s \tag{58}$$

where i is the frame distance between an initialized T-F unit and the target T-F unit.

The final reconstructed phase of the $(k,l)-th$ unit is obtained by integrating the frame-conditioned estimates with the weighted sum linear interpolation:

$$\hat{\phi}_{s}\left(k,l\right) = \frac{\sum_{i=-N_s}^{N_s}\left(s(i)\hat{M}\left(k,l+i\right)\right)\bar{\phi}_{i}\left(k,l\right)}{\sum_{i=-N_s}^{N_s}s(i)\hat{M}\left(k,l+i\right)} \tag{59}$$

where the $\bar{\phi}^i\left(k,l\right)$ is the smooth version of $\hat{\phi}^i\left(k,l\right)$, which is given by,

$$\bar{\phi}^i\left(k,l\right) = \text{unwrap}(\hat{\phi}^i\left(k,l\right) | \hat{\phi}^i\left(k,l-1\right)) \tag{60}$$

and $s(i)$ denotes the proximity weight for $\bar{\phi}^i\left(k,l\right)$, which is inversely proportional to the absolute value of the frame distance, that is, the phase estimate $\bar{\phi}^i\left(k,l\right)$ with a larger distance $|i|$ is assigned to a smaller proximity weight $s(i)$, and has a less effect to the final reconstructed phase. In our paper, we also set $s(i)$ to be the Hamming window as mentioned in [57].

3) Phase reconstruction with IFD: First of all, we calculate the estimated IF from the estimated normalized IFD.

Then, phase reconstruction with IF is similar to the process with GD. The only difference is that the former is reconstructed along with time axis, while the latter is along frequency axis. We generate the $(2N_s+1)$ frame-conditioned phase estimates for the $(k,l)-th$ T-F unit, which is given by:

$$\hat{\phi}^i\left(k,l\right) = \begin{cases} \hat{\phi}_{init}\left(k{+}i,l\right){+}\sum\limits_{n=0}^{i-1}\hat{\text{IF}}\left(k{+}n,l\right), \; i \neq 0 \\ \hat{\phi}_{init}\left(k{+}i,l\right), \qquad\qquad\qquad\qquad i = 0 \end{cases}, \qquad \forall \quad -N_s \leq i \leq N_s \tag{61}$$

with i denoting the frame distance.

Again, the phase reconstruction is considered as an integration of the frame-conditioned estimates using the weighted sum linear interpolation:

$$\hat{\phi}_s\left(k,l\right) = \frac{\sum_{i=-N_s}^{N_s}\left(s(i)\hat{M}\left(k{+}i,l\right)\right)\bar{\phi}_i\left(k,l\right)}{\sum_{i=-N_s}^{N_s}s(i)\hat{M}\left(k{+}i,l\right)} \tag{62}$$

with the smooth version $\hat{\phi}^i\left(k,l\right)$ given by

$$\bar{\phi}^i\left(k,l\right) = \text{unwrap}(\hat{\phi}^i\left(k,l\right) | \hat{\phi}^i\left(k-1,l\right)) \tag{63}$$

Unlike the work [57], which performs phase reconstruction with IFD only, our reconstruction is conducted with two phase derivative: IFD or GD. Finally, the estimated clean speech spectrogram can be obtained by

$$\hat{S}(k,l) = \left|\hat{S}(k,l))\right| e^{j\hat{\phi}_s(k,l)} \tag{64}$$

where the estimated magnitude is given by masking the noisy magnitude with the estimated mask $\hat{M}(k,l)$ from our DNN:

$$\left|\hat{S}(k,l)\right| = \hat{M}(k,l)\left|Y(k,l))\right| \tag{65}$$

The enhanced speech is then achieved by conducting the inverse STFT:

$$\hat{s}(n) = \mathrm{iSTFT}\{\hat{S}(k,l)\} \tag{66}$$

5.4 Experimental Results

5.4.1 Experimental Setup

The clean speech database used in our experiment is the TIMIT corpus [100], in which 731 utterances from different female and male speakers are used for the training and 87 utterances used for testing. Several types of noises are picked from the NOISEX-92 corpus [108], in which four types (babble, white, buccaneer1, factory) are regarded as seen noises, and the other four (pink, buccaneer2, street, hfchannel) as unseen noises. In the training stage, we mix the clean training speeches with seen noises at four levels (-3dB, 0dB, 3dB, 6dB) of signal-to-noise rates (SNRs) to obtain 11696 noisy speeches. In the enhancement stage, both seen noises and unseen noises are mixed with clean testing speeches at the above four SNR levels. The number of noisy utterances used in enhancement stage is 1392 for both seen noises and unseen noises. The sampling rate of all speech utterances and noises is set to 16 kHz. Hamming window is used in framing and the window size of STFT is 320 with 75$\%$ overlap. To assess the enhancement performance, three objective metrics are adopted in our experiment: PSEQ, STOI and SDR. For all metrics, a larger score indicates a better performance.

5.4.2 Selection of Controlling Factor

In this section, we investigate the enhancement performance when setting different values for the controlling factor μ. Firstly, we consider the following three types of lower and upper bounds for the local SNR as given in Table 8. Moreover, we set $\mu_{\min} = 1$, $\mu_{\max} = 10$ and $s = 25/(\mu_{\max} - \mu_{\min})$. A T-F unit will be treated as a noise dominant unit when the local SNR is under S_l. In contrast, a T-F unit is regarded as a speech dominant unit when the local SNR is above S_u.

Table 8: Different settings of lower and upper bounds

Type	S_l (dB)	S_u (dB)	μ_0
$\#1$	-15	10	$(3\mu_{\min} + 2\mu_{\max})/5$
$\#2$	-10	15	$(2\mu_{\min} + 3\mu_{\max})/5$
$\#3$	-5	20	$(\mu_{\min} + 4\mu_{\max})/5$
$\#4$	0	25	$\mu_{\max}$

As shown in Table 9, if S_l and S_u are very small, the noise dominant unit would be falsely classified to speech dominant unit, the residual noise would be fully removed and thus the scores of SDR and PESQ will decrease. On the contrary, if the S_l and S_u are too large, the speech unit with low local SNR will be mistakenly considered as noise, which could suppress the speech information leading to a decrease of STOI score. In terms of all metrics, the optimal case is type 3. The corresponding improvement of PESQ and SDR scores is significant, while the STOI score has no obvious degradation, which indicates that it removes the background noise quite well without extra speech distortion. For type 4, although it has the best PESQ and SDR scores, this setting is not perfect as the decrease of STOI score shows that the speech information is damaged compared with other settings. Secondly, we also investigate the different settings for $\mu_{\min}$ (varies from 0.5 to 1.5) and $\mu_{\max}$ (varies from 5 to 15). However, the objective results of the enhanced speech do not change significantly.

Table 9: Objective results with different controlling factors (on seen noise)

	PESQ				**STOI (%)**				**SDR**			
Methods	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB
Noisy	1.35	1.54	1.75	1.97	60.57	67.54	74.47	80.86	-2.85	0.11	3.08	6.07
Type1	1.98	2.25	2.51	2.78	**75.05**	**81.31**	**86.22**	**89.89**	7.08	9.29	11.33	13.46
Type2	2.00	2.27	2.53	2.79	74.82	81.18	86.08	89.76	7.29	9.49	11.52	13.62
Type3	2.01	2.29	2.55	**2.81**	74.79	81.02	86.01	89.61	7.44	9.63	11.64	13.73
Type4	**2.02**	**2.30**	**2.56**	**2.81**	74.71	80.95	85.91	89.57	**7.54**	**9.71**	**11.72**	**13.78**

5.4.3 Performance Comparison

Performance evaluation includes two main parts. First, we only consider magnitude-only masking techniques and compare our CRM with several existing masks. Second, we evaluate the complex spectrogram estimation methods by combining mask techniques with phase reconstruction. We adopted type 3 as the setting of the controlling factor. For fair comparison, all masks are estimated by the DNN with the same input features and configurations. As the difference between any two comparison methods is that the training targets of the DNN. Hence, we use the targets' names to represent each method in this section.

Table 10 illustrates the comparison masks as well as their definitions. Note that the first four masks (IRM, IAM, ORM and proposed CRM) only modify the magnitude spectrogram, while the cIRM enhances the complex spectrogram with Y_r and Y_i denote the RI parts of noisy speech, S_r and S_i the RI parts of clean speech, respectively.

Table 10: Comparison masks and definitions

Mask	Definition
IRM [36]	$\sqrt{P_s/(P_s+P_w)}$
IAM [126]	$\lvert S\rvert/\lvert Y\rvert$
OPM [138]	$(P_y+P_s-P_w)/2P_y$
CRM (proposed)	$P_s/(P_s+\mu P_w)$
cIRM [53]	$(Y_rS_r+Y_iS_i)/(Y_r^2+Y_i^2)+i(Y_rS_i-Y_iS_r)/(Y_r^2+Y_i^2)$

For the phase reconstruction based complex spectrogram estimation, we consider four methods, whose names are represented by mask + phase derivative. They are IRM+IFD [57], and proposed IRM+GD, CRM+IFD, CRM+GD. The mask is used for magnitude estimation, while the phase derivative is for phase reconstruction.

Seen noise

Table 11 gives the average objective score on seen noise. First, we compare proposed CRM with other traditional masking techniques. Obviously, the enhanced speech from CRM reaches the highest score under the metrics of PESQ and SDR as compared to IRM, IAM and OPM. More specifically, the proposed method has a large improvement on SDR scores, which means our enhanced speech has a higher SNR. The improvement indicates that our system strengthens the suppression of noise in noise dominant units using a large controlling factor. Our enhanced speech also obtains the best PESQ score, which reflects a good perceptual speech quality. The improvement of SDR and PESQ demonstrates that our CRM is better at noise suppression, especially in the noise dominant regions. Compared with the SDR and PESQ, the improvement of STOI is not that obvious. This is because the STOI algorithm mainly focuses on evaluating the intelligibility of the high-energy speech frames, and our CRM employs a small value of μ in speech dominant units. In this case, the value of our CRM is similar to those of other masks and thus the STOI scores are close for all tested methods. In terms of three metrics, we can conclude that our proposed CRM removes more residual noise while minimizing the speech distortion.

Furthermore, by comparing IRM with IRM+IFD and IRM+GD, we can clearly find that the methods with phase reconstruction show a considerable advantage over the magnitude-only method. Similar results can be found when comparing CRM with CRM+IFD and CRM+GD. However, the SDR score of CRM is better than the others in this group. Note that a multi-objective DNN is trained to estimate the mask and phase derivative simultaneously, with the assumption that these two training targets share similar structures. As CRM is a modified type of IRM, whose structure is less similar to the structure of phase derivative. Thus, the joint estimation of CRM and

Table 11: Objective scores of different methods on seen noise

	PESQ				**STOI (%)**				**SDR**			
Methods	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB
Noisy	1.35	1.54	1.75	1.97	60.57	67.54	74.47	80.86	−2.85	0.11	3.08	6.07
IRM	1.97	2.23	2.47	2.70	74.71	81.01	85.96	89.66	5.72	8.03	10.19	12.45
IAM	1.98	2.24	2.47	2.71	75.18	81.98	86.50	89.94	5.89	8.18	10.33	12.55
OPM	2.00	2.27	2.50	2.74	75.11	81.13	85.87	89.57	5.71	8.01	10.15	12.43
cIRM	**2.02**	2.28	2.54	2.78	77.07	83.08	87.69	91.03	6.71	8.97	11.01	13.07
IRM+IFD	2.01	2.27	2.52	2.75	76.50	82.34	86.69	90.35	6.08	8.39	10.53	12.68
IRM+GD	**2.02**	2.28	2.53	2.75	77.04	82.73	87.05	90.36	5.90	8.18	10.25	12.39
CRM	2.01	2.29	**2.55**	**2.81**	74.79	81.02	86.01	89.61	**7.44**	**9.63**	**11.64**	**13.73**
CRM+IFD	**2.02**	**2.30**	**2.55**	2.79	77.11	83.15	**87.72**	**91.04**	6.74	9.02	11.07	13.04
CRM+GD	**2.02**	2.29	2.54	2.78	**77.32**	**83.25**	87.67	90.92	6.42	8.65	10.65	12.60

the phase derivatives is not as accurate as the joint estimation of IRM and the phase derivatives. Therefore, although applying phase reconstruction to constrained masking technique has better objective scores in terms of PESQ and STOI, the improvement is not that significant and even suffers a decrease of SDR score. We can further understand the importance of complex spectrogram estimation by comparing the cIRM with the magnitude-only masks, whose objective scores are also better than the scores of IRM, IAM and OPM. Moreover, we notice that the improvement of STOI scores is the most significant when comparing the complex spectrogram estimation and magnitude-only enhancement, which indicates that the phase processing focuses more on increasing speech intelligibility.

Finally, we compare the performance of different complex spectrogram estimation methods. It should be pointed out that cIRM restores the complex spectrogram with the estimated RI spectrograms, while the mask+phase derivative method synthesizes the complex spectrogram with the estimated magnitude and reconstructed phase. The performance of cIRM is slightly better than IRM+IFD and IRM+GD, partly because IRM+IFD and IRM+GD still use noisy phase as the initial phase in the phase reconstruction. The CRM+IFD and CRM+GD achieve the best scores among all tested methods, which demonstrates the success of extending constrained masking to complex spectrogram.

Unseen noise

Table 12 shows the average objective scores on unseen noise. In general, our CRM+IFD still outperforms the other reference methods even under the unseen noise environment. However, compared to seen noise, the improvements of the objective scores on the enhanced speech decrease a bit, due to the increasing prediction error of masks. This result is not surprising since the mismatch in the types of the noises between the enhancement stage and the training stage makes the estimation of mask and phase derivative with DNN more difficult. Therefore, to improve the generalization capability under unmatched environments continues to be an important task for DNN-based methods. Comparing the results of complex spectrogram estimation methods on seen and unseen noises. We notice that the performance degradation of CRM+GD is larger than that of CRM+IFD. Similar results can be found with IRM+GD and IRM+IFD. This phenomenon indicates that the IFD is more robust under different noise conditions.

Table 12: Objective scores of different methods on unseen noise

	PESQ				**STOI (%)**				**SDR**			
Methods	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB
Noisy	1.34	1.52	1.73	1.94	62.61	69.46	76.20	82.69	−2.84	0.10	3.07	6.06
IRM	1.73	1.97	2.19	2.43	71.10	77.94	83.41	88.23	2.54	5.31	8.03	10.77
IAM	1.74	1.98	2.20	2.44	71.98	78.77	84.13	88.69	2.68	5.49	8.23	10.94
OPM	1.76	1.99	2.21	2.45	70.86	77.75	83.31	88.18	2.42	5.20	7.93	10.70
cIRM	1.77	2.00	2.23	2.44	71.99	79.58	85.12	89.24	3.61	6.42	8.82	11.46
IRM+IFD	1.78	2.02	2.23	2.47	72.77	79.47	84.73	89.20	2.92	5.73	8.41	11.04
IRM+GD	1.76	2.00	2.22	2.46	72.65	79.25	84.41	88.89	2.67	5.46	8.12	10.74
CRM	1.77	2.01	2.23	2.47	70.51	77.52	83.26	88.03	**3.92**	**6.72**	**9.39**	**12.01**
CRM+IFD	**1.80**	**2.06**	**2.28**	**2.52**	**74.20**	**81.02**	**86.07**	**90.14**	3.79	6.59	9.21	11.70
CRM+GD	1.78	2.02	2.25	2.49	73.73	80.35	85.38	89.57	3.26	6.05	8.66	11.19

5.4.4 Spectrograms of Enhanced Speeches

To better illustrate the benefits of constrained masking and phase reconstruction, the STFT spectrograms of the enhanced speeches from different methods are plotted and compared. The selected noisy speech is a mixture of clean speech and buccaneer noise at 0 dB. The whole STFT spectrogram is displayed in Fig. 24, from which we can observe the amount of the residual noise and the degree of speech distortion. Fig. 25 is obtained by zooming in the whole spectrogram and shows a part of the spectrogram with a time duration of 0.07 s to 0.15 s and a frequency range of 0 kHz to 4 kHz.

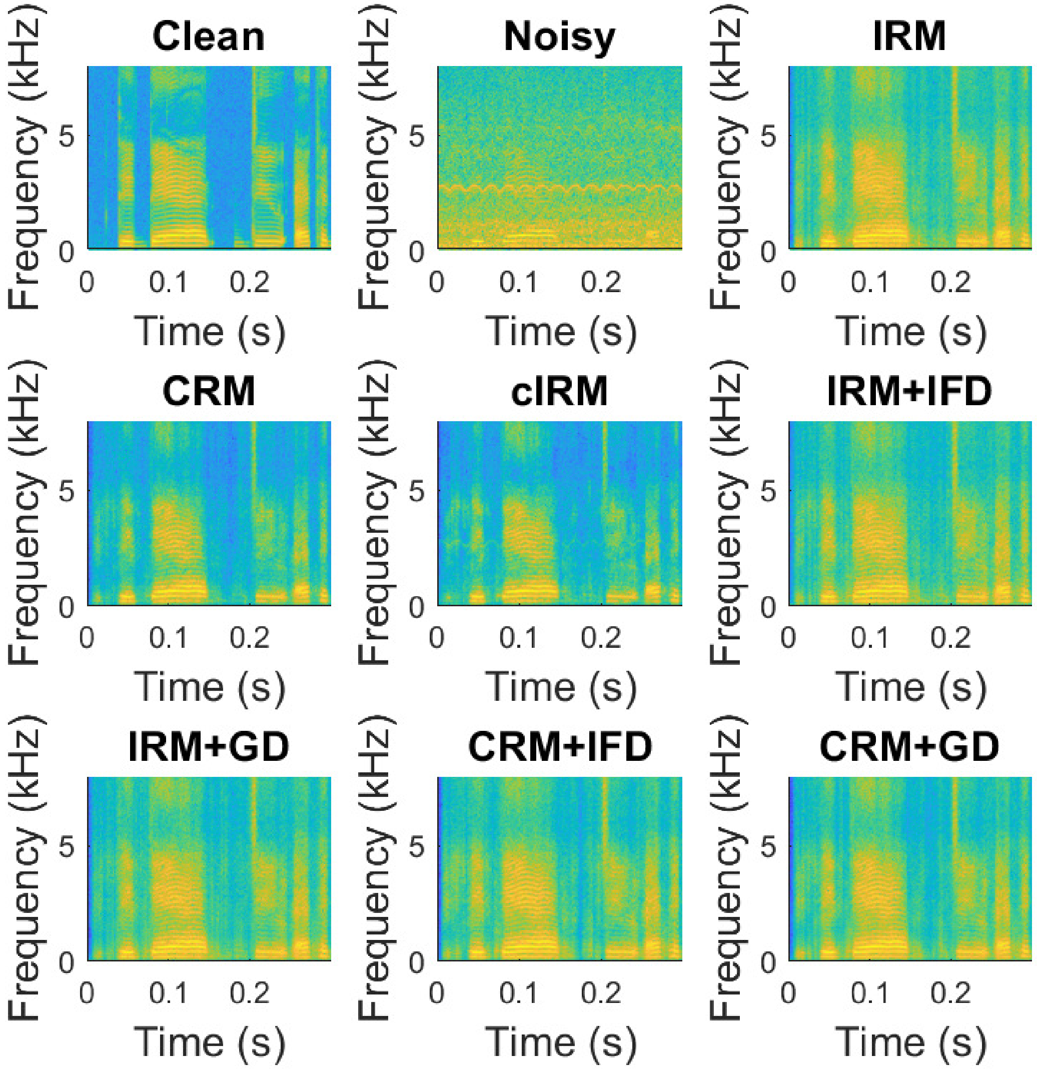


Figure 24: Spectrograms of the clean, noisy and enhanced speech signals for different methods (Whole spectrogram)

As Fig. 24 shows, the spectrogram of CRM is much clear than that of IRM, especially in the unvoiced frames. This indicates the enhanced speech from constrained masking technique contains less residual noise. Besides, the voiced frames of CRM-enhanced speech do not suffer notable distortion, which reveals the CRM's capability of controlling the trade-off between residual noise and speech distortion. We can also find that the amount of the residual noise in the spectrograms of CRM+IFD and CRM+GD is larger than that of CRM, this explains why CRM obtains better a SDR score than CRM+IFD and CRM+GD in Table 11 and 12.

Fig. 25 takes a closer look at the harmonic structures of the enhanced speeches. We focus on comparing the region in the red block of each enhanced speech and discover that the spectrograms with phase reconstruction (IRM+IFD, IRM+GD, CRM+IFD and CRM+GD) preserve better harmonic structures than the methods without phase reconstruction (IRM and CRM). Moreover, although the cIRM is also a method that jointly enhances the magnitude and phase spectra, its spectrogram is not as good as the spectrograms with phase reconstruction.

5.5 Conclusion

In this chapter, we have focused on improving the traditional T-F masking techniques for speech enhancement. First, we have proposed a new type of mask: CRM, by introducing a controlling factor to better adjust the trade-off between the speech distortion and residual noise in the enhanced speech. An FNN is then trained to estimate CRM from noisy feature set. The enhanced speech is synthesized with the enhanced magnitude processed by CRM and the noisy phase. Next, we have extended the CRM to the complex spectrogram estimation, where the enhanced magnitude spectrogram is obtained by applying the estimated CRM to the noisy magnitude spectrogram, and the estimated phase spectrogram is reconstructed with the noisy phase spectrogram and estimated phase derivatives. Performance evaluation demonstrates that the enhanced speech processed by the proposed CRM removes more background noise while does not suffer significant speech distortion as compared to existing mask techniques. Moreover, the performance has been further improved after extending CRM to complex spectrogram estimation. Objective results show that the speech

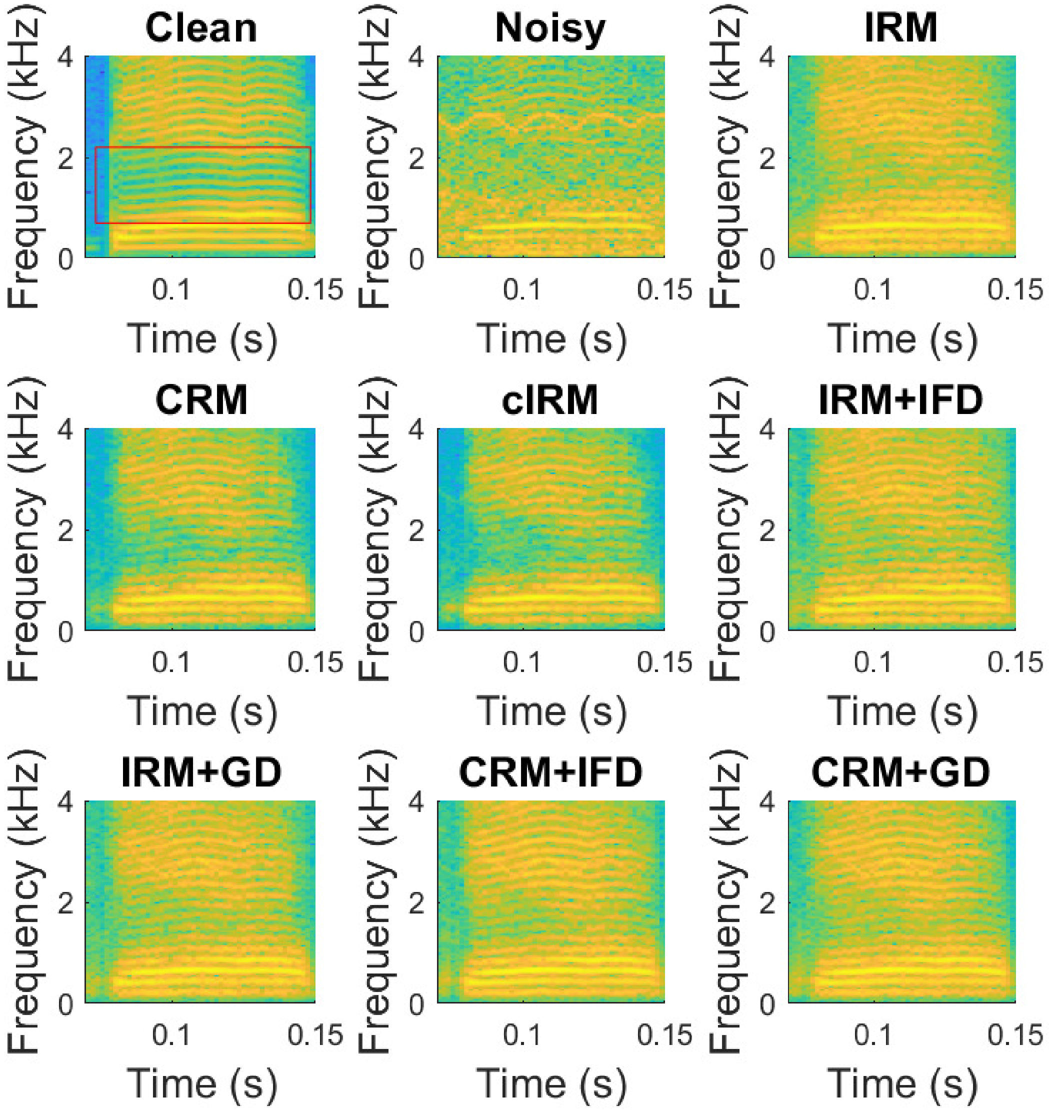


Figure 25: Spectrograms of the clean, noisy and enhanced speech signals for different methods (Local region)

intelligibility could be improved a lot with the phase reconstruction and the STFT spectrogram of the enhanced speech also exhibits a better harmonic structure.

Chapter 6

Conclusion and Future Work

6.1 Concluding Remarks

Thanks to its powerful learning capability, the DNN has been applied to SE in various ways and is still an ongoing area of research. In this thesis, we targeted the DNN-based joint enhancement of both magnitude and phase spectra from two aspects, i.e., time domain and T-F domain. In the time domain, we have developed several SE systems with DNN-augmented KF, whereas in the T-F domain, we aim at the modification of masking technique and then extend it to complex spectrogram estimation. Within each category, we proposed a few methods that resulted in performance improvements in various noisy conditions with respect to the most recent state-of-the-art variants.

In Chapter 3, we have presented a few novel systems based on basic KF. First, a DNN-augmented basic KF is proposed, where the DNN is adopted in parameter estimation to learn the LPCs variants: LSFs, from the feature set of the noisy speech. Then the basic KF with the estimated parameters is applied to the noisy speech for denoising. Experiments demonstrate that the DNN-augmented basic KF outperforms the traditional iterative KF. Moreover, by conducting extensive experiments under seen and unseen noises, we have observed that the DNN-augmented basic KF has a better generalization capability than the traditional DNN-based methods, as our system takes advantages of both supervised DNN-based technique and unsupervised statistical filtering. More specifically, the DNN-based parameter estimation offers more accurate parameters for Kalman filtering, whereas the denoising process with Kalman filtering is robust to different types of noises. Next, to further alleviate the degradation in HF component of the enhanced speech, we have implemented a DNN-based bandwidth extension procedure to the filtered speech for HF component restoration. Objective assessment shows improved speech quality resulting from the implementation of the restoration scheme. Finally, as an extension of the DNN-augmented basic KF, a hybrid system that combines DNN-based speech reconstruction and Kalman filtering is designed. The two-level processes in the hybrid system are performed in time and T-F domain, respectively, to make full use of the advantages in both domains. Performance assessment of the hybrid system reveals the improvement over the existing DNN-based methods.

Regarding the advanced KF in Chapter 4, we mainly focused on introducing DNN-based parameter estimation to two popular KFs: subband KF and colored-noise KF, respectively. The former divides the noisy speech into several subband speeches and performs Kalman filtering with each subband speech. The latter considers both clean speech and noise as AR processes, and removes the noise within colored-noise KF equations. In subband KF, an FNN is trained to estimate the LSFs of clean subband speech model. Whereas in colored-noise KF, both FNN and LSTM are used to estimate the LSFs of clean speech model and the LSFs of noise model simultaneously. A post subtraction is further applied to the colored-noise KF to remove the residual noise in the enhanced speech. Performance evaluation reveals that the proposed DNN-augmented advanced KFs outperform several traditional KF algorithms and existing DNN-based methods. Moreover, the DNN-augmented advanced KFs significantly alleviate the HF component degradation of the enhanced speech as compared to their counterpart: DNN-augmented basic KF.

In Chapter 5, the speech enhancement in the STFT domain using widespread T-F masking techniques is considered. First, we have derived a new type of mask: CRM, by introducing a spectral constraint to traditional T-F mask. The proposed mask aims to minimize the speech distortion while keeping the power of residual noise under a certain threshold. The CRM is estimated with an FNN trained on the noisy feature set and is then applied to the magnitude spectrogram for denoising. It is shown that the CRM better controls the trade-off between speech distortion and residual noise, thus achieves an improved performance in comparison to several existing masks. Next, we have extended the CRM to jointly enhance the magnitude and phase spectra, where the magnitude spectrogram is still denoised by the CRM, while the phase spectrogram is reconstructed using the noisy phase and the phase derivatives with the guide of CRM. Experiments demonstrate that the enhanced speech from the proposed complex spectrogram enhancement method has higher objective scores in terms of speech quality and intelligibility than that obtained by magnitude only enhancement.

6.2 Scope for the Further Work

Based on the literature review, technical contributions and experimental results made in this thesis, the following topics can be considered as prospective directions for future research.

- **Implementation with advanced DNN**: Most of the DNN structures in our proposed SE systems are the FNN or LSTM, which are not the most state-of-the-art ones. In Chapter 4, we have shown that LSTM outperforms FNN in the LSFs estimation for colored-noise KF, which proved the superiority of LSTM in modelling the temporal dependency of the sequence signals. However, one limitation of LSTM is the long processing time due to its expensive computational complexity. To improve the estimation accuracy of the DNN in our proposed DNN-based SE systems, more advanced DNNs, such as GAN or Transformer, could be adopted to substitute the FNN or LSTM. The GAN is able to generate more accurate data with the mini-max training scheme, while Transformer is a new structure which outperforms LSTM due to the reduced training time benefiting from parallel computation and the improved performance from processing long dependencies.

- **DNN-augmented modulation-domain KF**: The KF used in our SE system is time domain KF. Recently, several works have been proposed which apply the modulation-domain KF to

SE [26]. The modulation domain views the magnitude spectrum as a series of N modulating signals that span across time. Each modulating signal is then processed using a Kalman filter. The enhanced speech is obtained by synthesizing the processed magnitude spectrogram with the noisy phase spectrogram. Experiments in [26] demonstrate that the enhancement performance with modulation-domain KF is better than the one with time domain KF, if the ideal parameters are available for Kalman filtering. However, the system suffers significant performance degradation in practice, since the authors failed to provide accurate estimated parameters for KF with an iterative estimation algorithm. In our further work, we plan to apply the DNN-based parameter estimation into modulation-domain KF and compare its enhancement performance with the DNN-augmented time domain KF.

- **Perceptual ratio mask**: In Chapter 5, we have proposed the CRM to better control the trade-off between speech distortion and residual noise. However, the selection of the controlling factor is a difficult task in practice and does not take the perceptual properties into consideration. Inspired by the perceptual weighting filter in code-excited linear prediction (CELP) [148], which is originally used to shape the quantization noise by exploiting the masking properties of human ear, researchers have also proposed several algorithms to control the residual noise level by incorporating the perceptual weighting in speech enhancement [149, 150]. Therefore, we will propose a novel perceptual ratio mask for DNN-based speech enhancement. Unlike the traditional masks, which simply set large or small values for different T-F units according to the speech energy level in the noisy speech spectrum, our PRM makes use of the masking principle by incorporating a perceptual weighting filter in mask computation, which adaptively adjusts the value of PRM for each unit according to the perceptual weighting rules. As such, the enhanced speech is expected to keep the residual noise less audible while bringing no extra distortion to speech information.

- **Comparison between proposed time domain SE and T-F domain SE**: In this thesis, we have proposed two classes of methods incorporating phase information, that is, time-domain

KF-based SE systems and T-F domain masking based algorithms. The proposed methods show better performance over several SE algorithms existing in literature. However, the comparison between the two proposed classes has not been investigated. Therefore, our next work is to further investigate the time domain Kalman filtering with the T-F masking, in order to understand the benefits and limitations of each method and decide what method has overall minimum computational cost and best fits to a given specific application.

Bibliography

[1] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2007.

[2] J. Benesty and J. Chen, “Single-channel noise reduction with a filtering vector,” in *Optimal Time-Domain Noise Reduction Filters*, pp. 3–21, Springer, 2011.

[3] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, “Study of the Wiener filter for noise reduction,” in *Speech Enhancement*, pp. 9–41, Springer, 2005.

[4] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction Wiener filter,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[5] K. Paliwal and A. Basu, “A speech enhancement method based on Kalman filtering,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 12, pp. 177–180, 1987.

[6] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, “Raw waveform-based speech enhancement by fully convolutional networks,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 006–012, 2017.

[7] Y. Luo and N. Mesgarani, “Tasnet: time-domain audio separation network for real-time, single-channel speech separation,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, 2018.

[8] E. M. Grais, D. Ward, and M. D. Plumbley, “Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders,” in *European Signal Processing Conference (EUSIPCO)*, pp. 1577–1581, 2018.

[9] A. Pandey and D. Wang, “A new framework for CNN-based speech enhancement in the time domain,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 7, pp. 1179–1188, 2019.

[10] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6875–6879, 2019.

[11] A. Pandey and D. Wang, “Dense CNN with self-attention for time-domain speech enhancement,” *arXiv preprint arXiv:2009.01941*, 2020.

[12] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[13] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *Proc. of INTERSPEECH*, pp. 3642–3646, 2017.

[14] S. Qin and T. Jiang, "Improved wasserstein conditional generative adversarial network speech enhancement," *EURASIP Journal on Wireless Communications and Networking*, vol. 1, pp. 1–10, 2018.

[15] D. Baby and S. Verhulst, "SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 106–110, 2019.

[16] X. Hao, X. Su, Z. Wang, H. Zhang, *et al.*, "UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition," *arXiv preprint arXiv:2010.15521*, 2020.

[17] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[18] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5069–5073, 2018.

[19] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[20] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 208–211, 1979.

[21] W. M. Kushner, V. Goncharoff, C. Wu, V. Nguyen, and J. N. Damoulakis, "The effects of subtractive-type speech enhancement/noise reduction algorithms on parameter estimation for improved recognition and coding in high noise environments," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 211–214, 1989.

[22] L. Singh and S. Sridharan, "Speech enhancement using critical band spectral subtraction," in *Int. Conf. on Spoken Language Processing*, pp. 2827–2830, 1998.

[23] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4160–4164, 2002.

[24] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[25] T. Sreenivas and P. Kirnapure, "Codebook constrained wiener filtering for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 383–389, 1996.

[26] S. So and K. K. Paliwal, "Modulation-domain kalman filtering for single-channel speech enhancement," *Speech Communication*, vol. 53, no. 6, pp. 818–829, 2011.

[27] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[28] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP journal on Applied Signal Processing*, vol. 2005, pp. 1110–1126, 2005.

[29] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[30] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 15, no. 6, pp. 1741–1752, 2007.

[31] R. C. Hendriks, R. Heusdens, and J. Jensen, "Log-spectral magnitude mmse estimators under super-gaussian densities," in *Proc. of INTERSPEECH*, pp. 1319–1322, 2009.

[32] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[33] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[34] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[35] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[36] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7092–7096, 2013.

[37] Y. Li and S. Kang, "Deep neural network-based linear predictive parameter estimations for speech enhancement," *IET Signal Processing*, vol. 11, no. 4, pp. 469–476, 2016.

[38] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A deep neural network based harmonic noise model for speech enhancement," *Proc. of INTERSPEECH*, pp. 3224–3228, 2018.

[39] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, 2019.

[40] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5024–5028, 2018.

[41] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5039–5043, 2018.

[42] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[43] E. Loweimi, J. Barker, and T. Hain, "Statistical normalisation of phase-based feature representation for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5310–5314, 2017.

[44] G. Shi, M. M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 14, no. 5, pp. 1867–1874, 2006.

[45] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[46] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[47] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for stft spectrograms and their application to phase reconstruction.," in *Proc. of INTERSPEECH*, pp. 23–28, 2008.

[48] M. Krawczyk and T. Gerkmann, "Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1931–1940, 2014.

[49] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the stft-phase.," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129–132, 2013.

[50] F. Mayer, D. S. Williamson, P. Mowlaee, and D. Wang, "Impact of phase estimation on single-channel speech separation based on time-frequency masking," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4668–4679, 2017.

[51] J. Le Roux and E. Vincent, "Consistent wiener filtering for audio source separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 217–220, 2013.

[52] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 708–712, 2015.

[53] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 483–492, 2016.

[54] D. Williamson and D. Wang, “Time-frequency masking in the complex domain for speech dereverberation and denoising,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, 2017.

[55] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, “Complex spectrogram enhancement by convolutional neural network with multi-metrics learning,” in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2017.

[56] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, “A fully convolutional neural network for complex spectrogram processing in speech enhancement,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5756–5760, 2019.

[57] N. Zheng and X.-L. Zhang, “Phase-aware speech enhancement based on deep neural networks,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 1, pp. 63–76, 2018.

[58] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “PHASEN: A phase-and-harmonics-aware speech enhancement network.,” in *AAAI*, pp. 9458–9465, 2020.

[59] Y. Hu and P. C. Loizou, “Speech enhancement based on wavelet thresholding the multitaper spectrum,” *IEEE Trans. on Speech and Audio processing*, vol. 12, no. 1, pp. 59–67, 2004.

[60] P. J. Wolfe and S. J. Godsill, “Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 821–824, 2000.

[61] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, “On loss functions for supervised monaural time-domain speech enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 825–838, 2020.

[62] P. G. Shivakumar and P. G. Georgiou, “Perception optimized deep denoising autoencoders for speech enhancement.,” in *Proc. of INTERSPEECH*, pp. 3743–3747, 2016.

[63] Q. Liu, W. Wang, P. J. Jackson, and Y. Tang, “A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions,” in *European Signal Processing Conference (EUSIPCO)*, pp. 1270–1274, 2017.

[64] Y. Zhao, B. Xu, R. Giri, and T. Zhang, “Perceptually guided speech enhancement using deep neural networks,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5074–5078, 2018.

[65] Z. Zhao, S. Elshamy, and T. Fingscheidt, “A perceptual weighting filter loss for dnn training in speech enhancement,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 229–233, 2019.

[66] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1570–1584, 2018.

[67] H. Zhang, X. Zhang, and G. Gao, "Training supervised speech separation system to improve STOI and PESQ directly," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5374–5378, 2018.

[68] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 10, pp. 1702–1726, 2018.

[69] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1993–2002, 2014.

[70] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1593–1602, 1994.

[71] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[72] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[73] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[74] R. Vergin, D. O'shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, 1999.

[75] K.-H. Yuo and H.-C. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Communication*, vol. 28, no. 1, pp. 13–24, 1999.

[76] B. J. Shannon and K. K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1458–1485, 2006.

[77] S. Ikbal, H. Misra, and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 133–136, 2003.

[78] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4625–4628, 2009.

[79] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 7, pp. 1315–1329, 2016.

[80] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *Proc. of INTERSPEECH*, pp. 2058âĂŞ–2061, 2010.

[81] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[82] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 21, no. 2, pp. 270–279, 2013.

[83] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (pefac)," in *European Signal Processing Conference (EUSIPCO)*, pp. 451–455, 2011.

[84] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[85] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 5, pp. 1085–1094, 2017.

[86] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[87] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Proc. of INTERSPEECH*, pp. 3366–3370, 2013.

[88] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *IEEE Conf.on Computer Vision and Pattern Recognition (CVPR)*, pp. 1091–1100, 2018.

[89] D. Wang, "The time dimension for scene analysis," *IEEE Transactions on Neural Networks*, vol. 16, no. 6, pp. 1401–1426, 2005.

[90] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[91] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, 2013.

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

[93] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Int. Conf. on Machine Learning (ICML)*, pp. 1310–1318, 2013.

[94] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[95] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[96] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2013.

[97] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning (ICML)*, pp. 1139–1147, 2013.

[98] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[99] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[100] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report*, vol. 93, 1993.

[101] A. Rousseau, P. Deléglise, and Y. Esteve, "TED-LIUM: an automatic speech recognition dedicated corpus," in *Int. Conf. on Language Resources and Evaluation LREC*, pp. 125–129, 2012.

[102] A. Rousseau, P. Deléglise, Y. Esteve, *et al.*, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks.," in *Int. Conf. on Language Resources and Evaluation LREC*, pp. 3935–3939, 2014.

[103] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation," in *Int. Conf.on Speech and Computer*, pp. 198–208, 2018.

[104] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[105] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: an ASR corpus based on public domain audio books,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[106] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, “MLS: A large-scale multilingual dataset for speech research,” *Proc. of INTERSPEECH*, pp. 2757–2761, 2020.

[107] H.-G. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. of ISCA ITRW ASR*, pp. 181–188, 2000.

[108] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[109] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth ’CHiME’ speech separation and recognition challenge: dataset, task and baselines,” *arXiv preprint arXiv:1803.10609*, 2018.

[110] J. H. Hansen and B. L. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” in *Int. Conf. on Spoken Language Processing*, pp. 2819–2822, 1998.

[111] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[112] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, 2001.

[113] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4214–4217, 2010.

[114] J. D. Gibson, B. Koo, and S. D. Gray, “Filtering of colored noise for speech enhancement and coding,” *IEEE Trans. on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, 1991.

[115] S. Gannot, D. Burshtein, and E. Weinstein, “Iterative and sequential Kalman filter-based speech enhancement algorithms,” *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, 1998.

[116] T. Mellahi and R. Hamdi, “LPC-based formant enhancement method in Kalman filtering for speech enhancement,” *AEU-International Journal of Electronics and Communications*, vol. 69, no. 2, pp. 545–554, 2015.

[117] Y. Xia and J. Wang, “Low-dimensional recurrent neural network-based Kalman filter for speech enhancement,” *Neural Networks*, vol. 67, pp. 131–139, 2015.

[118] N. Nower, Y. Liu, and M. Unoki, “Restoration scheme of instantaneous amplitude and phase using Kalman filter with efficient linear prediction for speech enhancement,” *Speech Communication*, vol. 70, pp. 13–27, June, 2015.

[119] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, “Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 191–195, 2016.

[120] K. M. Jeon, S. Y. Park, C. J. Chun, N. I. Park, and H. K. Kim, “Multi-band approach to deep learning-based artificial stereo extension,” *ETRI Journal*, vol. 39, no. 3, pp. 398–405, 2017.

[121] I. V. McLoughlin, “Line spectral pairs,” *Signal Processing*, vol. 88, no. 3, pp. 448–467, 2008.

[122] M. H. Moattar and M. M. Homayounpour, “A simple but efficient real-time voice activity detection algorithm,” in *European Signal Processing Conference (EUSIPCO)*, pp. 2549–2553, 2009.

[123] W.-R. Wu and P.-C. Chen, “Subband Kalman filtering for speech enhancement,” *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 8, pp. 1072–1083, 1998.

[124] S. K. Roy, W.-P. Zhu, and B. Champagne, “Single channel speech enhancement using subband iterative Kalman filter,” in *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 762–765, May, 2016.

[125] K. Li and C.-H. Lee, “A deep neural network approach to speech bandwidth expansion,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4395–4399, 2015.

[126] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[127] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, “Veegan: Reducing mode collapse in GANs using implicit variational learning,” in *Advances in Neural Information Processing Systems*, pp. 3308–3318, 2017.

[128] D. C. Popescu and I. Zeljkovic, “Kalman filtering of colored noise for speech enhancement,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 997–1000, 1998.

[129] V. Grancharov, J. Samuelsson, and W. B. Kleijn, “Improved Kalman filtering for speech enhancement,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 1109–1112, 2005.

[130] N. Ma, M. Bouchard, and R. A. Goubran, “Perceptual Kalman filtering for speech enhancement in colored noise,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 717–720, 2004.

[131] N. Ma, M. Bouchard, and R. A. Goubran, “Speech enhancement using a masking threshold constrained Kalman filter and its heuristic implementations,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 19–32, 2005.

[132] S.-S. Wang, A. Chern, Y. Tsao, J.-w. Hung, X. Lu, Y.-H. Lai, and B. Su, “Wavelet speech enhancement based on nonnegative matrix factorization,” *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1101–1105, 2016.

[133] H. Yu, Z. Ouyang, W.-P. Zhu, and B. Champagne, “A deep neural network based Kalman filter for time domain speech enhancement,” in *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2019.

[134] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2005.

[135] T. Shimamura, N. Kunieda, and J. Suzuki, “A robust linear prediction method for noisy speech,” in *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, vol. 4, pp. 257–260, 1998.

[136] G. J. Brown and D. Wang, “Separation of speech by computational auditory scene analysis,” in *Speech Enhancement*, pp. 371–402, Springer, 2005.

[137] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, pp. 181–197, 2005.

[138] S. Liang, W. Liu, W. Jiang, and W. Xue, “The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio,” *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. EL452–EL458, 2013.

[139] K. Han and D. Wang, “A classification based approach to speech segregation,” *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[140] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2136–2147, 2015.

[141] S. Chakrabarty, D. Wang, and E. A. Habets, “Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks,” *IEEE Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 476–480, 2018.

[142] S. Gustafsson, P. Jax, and P. Vary, “A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics,” *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 397–400, 1998.

[143] P. C. Loizou, “Contrained Wiener filtering,” in *Speech enhancement: theory and practice*, CRC press, 2013.

[144] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[145] A. P. Stark and K. K. Paliwal, “Speech analysis using instantaneous frequency deviation,” in *Proc. of INTERSPEECH*, 2008.

[146] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, “Significance of the modified group delay feature in speech recognition,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[147] V. K. Prasad, T. Nagarajan, and H. A. Murthy, “Automatic segmentation of continuous speech using minimum phase group delay functions,” *Speech Communication*, vol. 42, no. 3-4, pp. 429–446, 2004.

[148] P. Kroon and B. S. Atal, “Predictive coding of speech using analysis-by-synthesis techniques,” in *Asilomar Conf. on Signals, Systems and Computers*, vol. 2, pp. 664–664, 1990.

[149] Y. Hu and P. C. Loizou, “A perceptually motivated approach for speech enhancement,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 457–465, 2003.

[150] M. F. A. Chowdhury, M. J. Alam, M. F. Alam, and D. O’Shaughnessy, “Perceptually weighted multi-band spectral subtraction speech enhancement technique,” in *IEEE Int. Conf. on Electrical and Computer Engineering*, pp. 395–399, 2008.