Artificial Bandwidth Extension Using H^∞ Sampled-data Control Theory and Speech Production Model

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Deepika Gupta

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

March 2022

Certificate

This is to certify that the thesis entitled "**Artificial Bandwidth Extension Using H^∞ Sampled-data Control Theory and Speech Production Model**", submitted by **Deepika Gupta** (156102023), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by her under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:
Guwahati.

Dr. Hanumant Singh Shekhawat
Assistant Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

My parents and family

who is everything in my life

My husband **Alok Mittal**

for his love and sacrifice

My guide **Dr. Hanumant Singh Shekhawat**

for his guidance and inspiration

My **parents** and **parents-in-law**

for their blessings

My **brothers** and **sisters in-law**

for their love and support

Acknowledgements

This thesis would not have been possible without the immense help and support of several people in various measures. I would like to convey my acknowledgment to all of them.

First and foremost, I express my sincere gratitude to my research supervisor, Dr. Hanumant Singh Shekhawat for his continuous guidance in all aspects and constant motivation throughout the doctoral studies. His insightful feedbacks have helped greatly in improving the quality of my thesis.

I am very thankful to former chairmen Prof. S.R.M. Prasanna and current chairmen Prof. Rohit Sinha of my doctoral committee for investigating my work and for guiding me in right direction. I want to express my sincere gratitude to the other doctoral committee members, Prof. M. K. Bhuyan and Dr. Tony Jacob for their valuable suggestions on my work and for spending their valuable time in evaluation of my work.

I would like to say thank to other teaching and non-teaching staffs of the EEE department for their care, help and support throughout my Ph.D. duration; especially, I acknowledge Mukut Sir for timely forwarding of different applications.

This thesis would become highly impossible without the help of past/present members of the department Shikha, Mrinmoy, Shoubhik, Protima, Akhilesh, Vikram, Brij Nandan, Debajit, Balaji, Sandeep, Sarfaraz, Sreeram, Moa, Pradipta, Anik, Sishir, Abhimanyu, Vineeta, Tilendra, Sibasis, Sukanya and the rest for their help in subjective evaluation of the work. I never forget my friends Shikha and Mrinmoy for their support and help.

I want to thank the funding agencies, such as IIT Guwahati and IndSCA for funding travel to attend conferences abroad, which allowed me to meet experienced researchers and explore very nice destinations like Graz and Vienna. I would like to thank MHRD, Government of India, for providing me a fellow-ship to pursue my Ph.D.

My deepest gratitude goes to my husband. Without his love, support, and sacrifice it wouldn't have been possible for me to complete my PhD.

Deepika Gupta

Abstract

This thesis aims to enhance the quality of the narrowband speech signal transmitted in narrowband telephonic communication. The transmitted narrowband speech signal has frequency components in the range of 300-3400 Hz. Original speech signal consists of significant frequency components beyond this limit, making it easier to understand the speech signal, i.e., the speech quality and intelligibility are improved. Therefore, the received narrowband signal at the receiver end in the narrowband telephonic communication can be enhanced by recovering missing high-frequency components in the speech signal, typically in the frequency range 4-8 kHz. A process of recovering high-frequency components is known as an artificial bandwidth extension (ABE) process. The ABE process improves speech intelligibility and quality. The thesis proposes artificial bandwidth extension frameworks using the H^∞ sampled-data control theory and machine learning techniques. The performances of the proposed approaches have been evaluated by using objective and subjective measures. Also, these measures are computed for the two different datasets.

Keywords: H^∞ sampled-data control theory, bandwidth extension, speech production model, deep neural network modeling, modulation.

Contents

List of Figures xvii

List of Tables xxi

List of Acronyms xxv

1 Introduction 1

1.1 Artificial bandwidth extension . . . 2

1.2 Review of current ABE approaches . . . 4

1.3 Motivation, challenges, and our aims . . . 10

1.4 Contributions of the Thesis . . . 13

2 A new paradigm in artificial bandwidth extension 17

2.1 A Proposed set-up based on wideband modeling for artificial bandwidth extension of speech signals . . . 19

2.1.1 Training block . . . 19

2.1.1.1 Windowing and framing . . . 20

2.1.1.2 Wideband feature vector extraction . . . 20

2.1.1.3 Computation of $F(z)$ . . . 23

2.1.1.4 Narrowband feature vector extraction . . . 24

2.1.1.5 Modeling . . . 24

2.1.2 Extension block . . . 27

2.1.2.1 Wideband feature vector estimation . . . 27

2.1.2.2 Wideband signal estimation . . . 27

2.2 Experimental analysis and results . . . 28

2.2.1 Databases 28

2.2.2 Objective analysis 28

2.2.2.1 Performance evaluation using Gaussian mixture model 30

2.2.2.2 Performance evaluation using deep neural network 30

2.2.2.3 Performance comparison 33

2.2.3 Subjective listening test 35

2.3 Conclusion 36

3 Artificial bandwidth extension technique based on the wideband modeling 37

3.1 A proposed set-up based on wideband modeling for artificial bandwidth extension of speech signals 39

3.1.1 Designing of the pre-trained model 39

3.1.1.1 Pre-processing of speech signals 40

3.1.1.2 High-band feature extraction 41

3.1.1.3 Gain calculation 45

3.1.1.4 Narrowband feature vector extraction 45

3.1.1.5 Training of the DNN model 46

3.1.2 Artificial bandwidth extension of AMR coded narrowband speech signal . 46

3.1.2.1 Narrowband signal reconstruction process 47

3.1.2.2 High-band feature vector and gain factor Estimation 47

3.1.2.3 High-band signal estimation 47

3.1.2.4 Wideband signal estimation 47

3.2 Experimental set-up and results 48

3.2.1 Databases 49

3.2.2 Results 49

3.2.2.1 Architecture of the DNN model 50

3.2.2.2 Objective assessment 52

3.2.2.3 Subjective assessment 53

3.3 Conclusion 54

4 Artificial bandwidth extension technique based on the high-band modeling 59

4.1 A proposed set-up based on high-band modeling for artificial bandwidth extension of speech signals . . . 61

4.1.1 Training block . . . 61

4.1.1.1 Framing . . . 62

4.1.1.2 High-band feature vector extraction . . . 62

4.1.1.3 Gain factor calculation . . . 66

4.1.1.4 Narrowband feature vector extraction . . . 66

4.1.1.5 Modeling . . . 67

4.1.2 Extension block . . . 67

4.1.2.1 Narrowband signal process . . . 67

4.1.2.2 Mapping process . . . 68

4.1.2.3 Estimation of the high-band signal . . . 68

4.1.2.4 Wideband signal estimation using the DFT concatenation . . . 68

4.2 Experiment analysis and results . . . 69

4.2.1 Databases and parameters . . . 69

4.2.2 Objective analysis . . . 70

4.2.3 DNN model performance . . . 71

4.2.4 Performances comparison . . . 73

4.2.5 Subjective listening test . . . 75

4.3 Conclusion . . . 75

5 Artificial bandwidth extension technique based on the mapped high-band modeling 77

5.1 Proposed framework for the artificial bandwidth extension of speech signal . . . 78

5.1.1 Designing of the pre-trained models . . . 79

5.1.1.1 Speech file operations . . . 80

5.1.1.2 Band pass shifted feature vector extraction . . . 82

5.1.1.3 Narrowband feature vector extraction . . . 85

5.1.1.4 Designing of DNN-1 model 86

5.1.1.5 Gain factor computation 86

5.1.1.6 Designing of DNN-2 model 87

5.1.2 Extension of the AMR coded narrowband signal 87

5.1.2.1 Narrowband signal reconstruction process 87

5.1.2.2 Band pass shifted feature vector prediction 88

5.1.2.3 Gain factor prediction 88

5.1.2.4 High-band signal estimation 88

5.1.2.5 Wideband signal synthesis 88

5.2 Speech databases, measures, and results analysis 89

5.2.1 Databases 89

5.2.2 Measures for performance evaluation 90

5.2.3 Results analysis 90

5.2.3.1 DNN-1 model architecture 90

5.2.3.2 DNN-2 model architecture 92

5.2.3.3 Objective comparison with baselines 93

5.2.3.4 Subjective comparison 94

5.3 Objective comparison with the previous schemes 95

5.3.1 An objective comparison in the oracle conditions 96

5.3.2 An objective comparison in practical conditions 97

5.4 Conclusion 97

6 Summary and future work 101

6.1 Summary of the work 102

6.2 Future directions 104

A Sampled-data system theory 107

A.1 History of sampled-data system theory 108

A.2 Abbreviations 109

A.3 A general closed-loop system 109

A.4 Lifting and inverse lifting 110

B A general solution of sampled-data system problem in ABE 113

C Objective measures 117

Bibliography 121

List of Publications 129

List of Figures

1.1 A general block diagram of the artificial bandwidth extension technique used at the receiver side. 3

1.2 A general architecture of the error system. 11

2.1 Block diagram consists of training of a model and extension of the narrowband signal. 20

2.2 Error system set-up for reconstructing of a stationary speech signal. 20

2.3 Proposed architecture of error system with considering signal modeling for reconstructing a stationary speech signal. 21

2.4 Spectrogram of **(a)** Original wideband signal, **(b)**, **(c)**, and **(d)** reconstructed wideband signal by the proposed method, spectral translation, cepstral domain, respectively. 35

3.1 Block diagram Illustrating the training of the DNN model. 40

3.2 AMR coded narrowband signal generation process. 40

3.3 Wideband signal generation process. 41

3.4 A proposed error system for wideband signal reconstruction. 41

3.5 Proposed an error system with pole-zero modeling for wideband signal reconstruction. 42

3.6 Bandwidth extension technique for the AMR coded narrowband signal. 45

3.7 Illustration of the artificial bandwidth extension of the AMR coded narrowband signal. 46

3.8 Spectrogram of **(a)** reference wideband speech signal of a female speaker, **(b)** AMR coded narrowband signal $S_{AMR-NB}[n']$ sampled at 16 kHz, (**c, d, and e**) extended wideband speech signals using the signals $\tilde{S}_{BPF}[n']$, $\widehat{S}_{HB}[n']$, and $10^{\frac{d}{20}}\widehat{S}_{HB}[n']$ in the proposed framework using DNN model (see Figure 3.7). . . . 56

3.9 Spectrogram of **(a)** reference wideband speech signal of a female speaker, **(b)** AMR coded narrowband signal sampled at 16 kHz, and **(c,d,e)** extended speech signals by the proposed approach, modulation technique, and cepstral domain approach, respectively 57

4.1 Block diagram consists of training of DNN model and artificial bandwidth extension of the narrowband signal. 62

4.2 Bandwidth extension process applied to a stationary narrowband signal in order to estimate the corresponding high-band signal. 62

4.3 An error system set-up. 63

4.4 A proposed architecture of the error system set-up for estimating the high-band signal. 64

4.5 Spectrogram of: **(a)** artificially extended speech signal by the cepstral domain approach using DNN model, **(b)** artificially extended speech signal by the **modulation technique** using DNN model, **(c)** artificially extended speech signal by the **proposed approach** using DNN model, and **(d)** original WB signal 74

5.1 Illustrating the training of Deep Neural Networks. 79

5.2 AMR coded narrowband signal production process. 80

5.3 Wideband signal production process. 81

5.4 Band pass shifted signal production process. 82

5.5 A proposed error system. 82

5.6 A proposed error system considers the signal modeling. 83

5.7 Estimation of the band pass filtered signal. 86

5.8 Illustrating the artificial bandwidth extension of the coded narrowband signal. . 87

5.9 Spectrogram of **(a)** reference wideband speech signal of a female speaker, **(b)** AMR coded narrowband signal sampled at 16 kHz, and **(c,d,e)** extended speech signal by the proposed approach, modulation technique, and cepstral domain approach, respectively 99

A.1 Single rate discrete-time lifted system [1] 110

B.1 A general sampled-data error system in ABE. 114

B.2 General standard feedback control system. 115

List of Tables

2.1 Performance comparison of extended speech signals enhanced by applying the upsampler (without applying filter K) and the oracle IIR interpolation filter K in Figure 2.1 on the speech files taken from the validation set. 29

2.2 Performance evaluation for some speech files taken from the validation set in condition of direct implanting FIR filter K (oracle K) in Figure 2.1 for ABE. . 29

2.3 Performance evaluation by using 128 GMMs on the test set. 30

2.4 Performance evaluation on the validation set for different DNN topologies by varying the number of hidden layers (N_{HL}) and the number of units (N_U), and ReLU activation function in hidden layers, linear activation function in the output layers, LSF narrowband features and *AdaMax* optimizer. 31

2.5 Performance evaluation on the validation set for the DNN model designed using 4 hidden layers and 256 units in each hidden layer for different batch sizes. . . . 31

2.6 Performance evaluation on the test set by the proposed approach using the DNN model designed using 4 hidden layers and 256 units in each hidden layer for different narrowband features. 32

2.7 Performance evaluation for the voiced speech extracted from speech files belonging to the test set using the DNN model designed using 4 hidden layers and 256 units in each hidden layer for different narrowband features. 32

2.8 Performance evaluation for the unvoiced speech extracted from speech files belonging to the test set using the DNN model designed using 4 hidden layers and 256 units in each hidden layer for different narrowband features 33

2.9 A comparison of the objective measures computed on the test set speech files for the proposed approach and the baselines 34

2.10 Subjective assessment on artificially extended speech files belonging to the test set by the proposed method with respect to the baselines 36

3.1 Performance evaluation of enhanced speech files belonging to the validation set in the condition of directly using the FIR synthesis filter obtained by truncating the impulse response of IIR synthesis K_{HPF} and without applying the SFS technique (d = 0) for ABE 49

3.2 Performance evaluation of enhanced speech files belonging to the validation set in the condition of directly using the FIR synthesis filter obtained by truncating the impulse response of IIR synthesis K_{HPF} and applying the SFS technique (d $\neq$ 0) for ABE 50

3.3 Computation of the mean squared error and standard deviation for the validation set with varying the DNN architecture 51

3.4 Performance evaluation on the test set for the proposed approach and the baselines. 52

3.5 Subjective assessment conducted on the artificially extended speech files belonging to the test set. 54

4.1 Performance evaluation on the validation set in condition of direct use of FIR filter obtained by truncating the impulse response of IIR K_{opt} in Figure 4.1 for ABE. 70

4.2 Performance evaluation on the validation set for the signals $S_{NB}[n']$, $S'_{NB}[n'] = g_3 S_{NB}[n']$, $S'_{WB}[n']$ in Figure 4.1 for ABE. 71

4.3 Objective analysis on the validation set by varying the learning rate and mini-batch size for the fixed DNN topology with 3 N_{HL} and 512 N_U and Relu activation function in hidden layers. 72

4.4 Objective analysis on the validation set by varying the number of hidden layers (N_{HL}) and the number of units (N_U) in hidden layer for the fixed batch size 768, and Relu activation function in hidden layers. 72
4.5 Objective analysis on the test set for the proposed approach and the existing approaches. 73

5.1 Computation of wideband MOS-LQO for the validation set with varying the DNN architecture 91
5.2 Wideband MOS-LQO computation for the extended speech signals of the validation set using the DNN architecture designed with 2 hidden layers and 64 neurons. 91
5.3 Computation of wideband MOS-LQO for the validation set with varying the DNN architecture 92
5.4 Performance evaluation on the test set for the proposed approach and baselines. 93
5.5 Subjective assessment conducted on the artificially extended speech files belonging to the test set. 95
5.6 Performance evaluation of each modeling scheme for the speech files belonging to the validation set. 96
5.7 Performance comparison between wideband modeling and mapped high-band modeling in practical scenario. 97

List of Acronyms

ABE	Artificial bandwidth extension
NB	Narrowband
WB	Wideband
HB	High-band
LP	Linear prediction
BP-MGN	Bandpass-envelope modulated Gaussian noise
LSF	Line spectral frequencies
MFCC	Mel frequency cepstral coefficients
LPC	Linear prediction coefficients
Cepstrum	Linear frequency cepstral coefficients
VQ	Vector quantization
GMM	Gaussian mixture model
HMM	Hidden Markov model
DNN	Deep neural network
SFS	Spectral floor suppression
AM-FM	Amplitude modulation and frequency modulation
AMR	Adaptive multi rate
LTI	Linear time-invariant
DFT	Discrete Fourier transform
pdf	Probability distribution function
MMSE	Minimum mean squared error
AdaMax	Adaptive moment estimation based on the infinity norm

MVN	Mean and variance normalization
LPF	Low pass filter
HPF	High pass filter
BPF	Band pass filter
MSE	Mean square error
SDR	Signal to distortion ratio
LLR	Log likelihood ratio
LSD	Logarithmic spectral distance
MOS-LQO	Mean opinion score listening quality objective
IIR	Infinite impulse response
FIR	Finite impulse response
MSIN	Mobile station input
HQ2	High-quality low pass filter
PESQ	Perceptual evaluation of speech quality
CMOS	Comparison mean opinion score
LDTI	Linear discrete time-invariant

1

Introduction

Contents

1.1 Artificial bandwidth extension 2

1.2 Review of current ABE approaches 4

1.3 Motivation, challenges, and our aims 10

1.4 Contributions of the Thesis 13

Objective

Artificial bandwidth extension (ABE) is an enhancement technique, which extends the bandwidth of a signal. ABE technique is used in narrowband (NB) telephonic communication to create new frequency components in the band from 3.4 kHz (or 4 kHz) to 7 kHz (or 8 kHz). In narrowband telephonic communication, the bandwidth of the transmitted narrowband speech signal sampled at 8 kHz is limited to 300-3400 Hz [2]. As a result, quality and intelligibility of the transmitted narrowband speech signal degrade. Therefore, the artificial bandwidth extension technique can be applied to the received narrowband signal at the receiver side for artificially regenerating the missing high-frequency components. It improves the perception of the speech sounds. At present, the wideband (WB) speech services are provided, which transmit the wideband signal sampled at 16 kHz. The transmitted wideband signal consists of frequency components in the range of 50-7000 Hz. As a result, the speech quality and intelligibility of the received wideband signal are perceived better. But, wideband speech services require up-gradation of the terminal devices at both the ends (transmitting and receiving) and transmission network. It is time-consuming and costly. Therefore, narrowband speech services are used. ABE technique can be utilized in narrowband speech communication systems without modifying the existing transmitter set-up and transmission network.

1.1 Artificial bandwidth extension

A general process is shown in Figure 1.1 for ABE. In Figure 1.1, $\uparrow 2$ denotes an upsampler with the upsampling factor 2, which is defined as

$$u[n] = \begin{cases} v[n/2], & n = 0, 2, 4, 6, ... \\ 0, & \text{otherwise} \end{cases} \tag{1.1}$$

where v and u are discrete input and output, respectively. Estimation of the wideband signal is performed in two main parallel processes at the receiver side. One is the high-band signal (at 16 kHz) reconstruction, and another is the narrowband signal (at 16 kHz) reconstruction, as shown in Figure 1.1. The high-band (HB) signal reconstruction process includes narrowband features

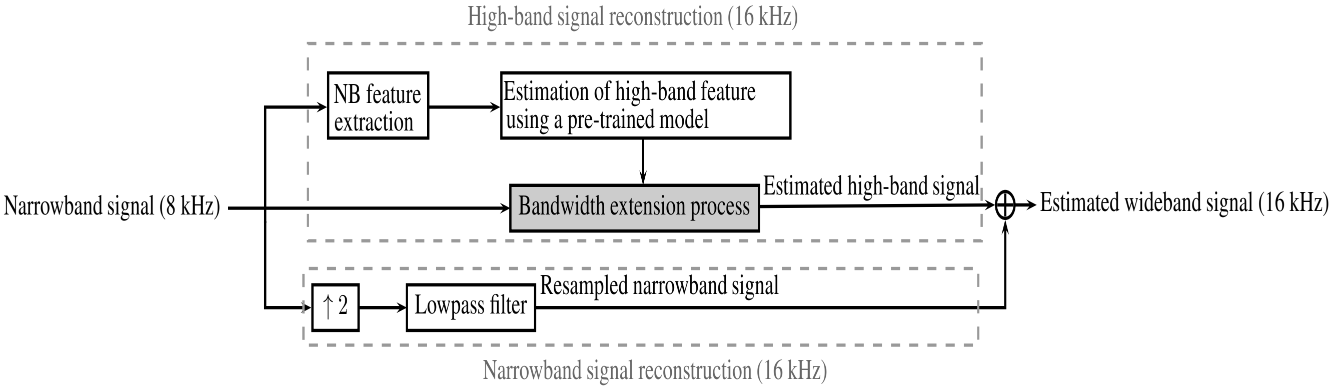
High-band signal reconstruction (16 kHz)
NB feature extraction
Estimation of high-band feature using a pre-trained model
Narrowband signal (8 kHz)
Bandwidth extension process
Estimated high-band signal
Estimated wideband signal (16 kHz)
↑2
Lowpass filter
Resampled narrowband signal
Narrowband signal reconstruction (16 kHz)

Figure 1.1: A general block diagram of the artificial bandwidth extension technique used at the receiver side.

(narrowband information) extraction, high-band features (high-band information) estimation using a pre-trained model, and bandwidth extension process. The bandwidth extension process needs the high-band features for estimating the high-band signal. The high-band features have attributes of the high-band signal, which can not be obtained the same for all speech sounds because of the non-stationary (time-varying) behavior of speech sounds [3]. This leads to the need of numerous high-band features for reconstructing the full speech signal. Therefore, the high-band features are estimated using modeling techniques. The fundamental idea is to associate the narrowband features with the high-band features using a model designed by machine learning technique. The model is trained using the narrowband features and corresponding high-band features which we call the pre-trained model. The model training is an offline process. The non-stationary nature of speech sounds makes this high-band feature estimation process a little challenging too. The narrowband features are computed using the narrowband signal, which has characteristics of the narrowband signal. Only narrowband information is available on the receiver side; hence the pre-trained model helps in the estimation of the high-band signal even though the high-band information is missing on the receiver side. The estimated high-band features are used in the bandwidth extension process, which synthesizes the high-band signal. The narrowband signal reconstruction process is relatively easy and pretty standard. The narrowband signal (8 kHz) at the receiver end is resampled at 16 kHz. To this end, the narrowband signal is passed through the upsampler (↑ 2), followed by a lowpass filter. The upsampling of the narrowband signal produces an unwanted mirror image of the narrowband spectrum in the high-band region at the output. Hence, it is removed by the lowpass filter. In the end, the wideband signal is reconstructed by adding the estimated high-band signal and the resampled narrowband signal. Different ABE techniques explained in following Section 1.2 mainly differ in narrowband features, high-band features, machine learning modeling techniques, and bandwidth extension process.

1.2 Review of current ABE approaches

Many ABE approaches have been developed in which most of them are based upon the speech production model (source-filter model) for speech production [4]. In the speech production model, the speech signal is segmented into a speech production filter and a residue signal/ excitation signal. A speech signal is an output of the speech production filter driven by the excitation signal. The speech production filter models the combined effect of the vocal tract and the radiation at the lips, as well as the glottal pulse shape in the case of voiced sounds. The excitation signal can be a white noise for unvoiced speech, a quasi-periodic impulse train for voiced speech, or a combination of them. In both the cases, the magnitude spectrum of the excitation signal is flat. Thus, the speech production filter consists of the spectral envelope of the speech signal. Most of the ABE approaches typically use an all-pole model (autoregressive model) to represent the speech production filter. The speech production filter and excitation signal can be obtained using a linear prediction (LP) method [5, 6]. LP model has two main processes: LP analysis and LP synthesis. In the LP analysis, the speech signal is decomposed into the speech production filter and excitation signal using an LP analysis filter. In the LP synthesis, the speech signal is reconstructed by passing the excitation signal through the speech production filter (LP synthesis filter). The LP analysis filter is an inverse form of the LP synthesis filter.

In ABE methods based on the speech production model, the high-band spectral envelope and the high-band excitation of the wideband signal are estimated. The high-band excitation can be estimated directly using the narrowband excitation. For this, the narrowband excitation is processed by a residual extension method. Several residual extension methods are developed, such as spectral folding [7–9], spectral translation [7, 9–13], pitch adaptive modulation [9, 11], bandpass-envelope modulated Gaussian noise (BP-MGN) [9, 14], and full-wave rectification [7, 9, 15], which are explained as follows.

- In the spectral folding method, the narrowband excitation signal is up-sampled by a factor of 2 for generating the high-band excitation signal. This method causes the spectral gap around 4 kHz and does not preserve the harmonic structure in high-band.
- In the spectral translation method, the spectrum of the narrowband excitation signal is shifted by a fixed modulation frequency, which yields the high-band excitation signal. It can fill the spectral gap around 4 kHz by choosing the appropriate modulation frequency but does not preserve the harmonic structure in high-band.
- In the pitch adaptive modulation method, the modulation frequency is adapted and chosen in such a way that it is an integer multiple of the fundamental frequency of speech (pitch). This method needs an accurate detection of the fundamental frequency. This method preserves the harmonic structure in high-band but is sensitive to a small error in pitch detection.
- In the full-wave rectification method, the high-band excitation is obtained by rectifying the narrowband excitation sampled at 16 kHz. It maintains the harmonic structure but needs to control the energy level of the synthesized excitation in high-band.
- In the bandpass-envelope modulated Gaussian noise (BP-MGN) method, the high-band excitation is generated by modulating the bandpass-envelope with Gaussian noise. The bandpass-envelope is extracted from the narrowband signal sampled at 16 kHz.

The high-band spectral envelope is varied for different speech sounds/phonemes because of the time-varying behavior of speech sounds [3]. Therefore, it is estimated using the pre-trained model. The design process of the pre-trained model requires high-band information (high-band features) and corresponding narrowband information (narrowband features). These features can represent spectral envelope information. The high-band spectral envelope and the narrowband spectral envelope can be represented by the line spectral frequencies (LSF) [8,14,16], Mel frequency cepstral coefficients (MFCC) [8], linear prediction coefficients (LPC) [10,12,17], and linear frequency cepstral coefficients (cepstrum) [11, 13, 15] features. The high-band information for given narrowband information is estimated using the pre-trained model. This model is designed using machine learning techniques, for example, linear mapping approach [18], codebook mapping approach like vector quantization (VQ) [10, 12, 19], and statistical modeling approaches like Gaussian mixture models (GMMs) [20–24], hidden Markov models (HMMs) with GMMs [13, 25–28], and deep neural network (DNN) topologies [13, 16, 29–32]. ABE approaches based on the speech production model have been developed using the combination of residual extension method, spectral envelope representation, and spectral envelope estimation method.

In [8], the bandwidth extension is implemented using the spectrum folding excitation extension method, MFCC features for the narrowband spectral envelope, LSF features for the wideband spectral envelope, and VQ codebook approach. While in [18], both the narrowband and high-band information are represented by the LSF features, and the linear mapping function is used to estimate the high-band LSF features. In linear mapping, four mapping matrices are used for a better prediction of the high-band LSF features. These mapping matrices are clustered using first two reflection coefficients of the narrowband speech signal [5].

In [19], the ABE framework consists of the bandpass-envelope modulated Gaussian noise for excitation extension, the LSF features and lowpass energy prediction error for the narrowband features, the LSF features and high-band gain for the high-band features, and the VQ codebook approach. This ABE scheme focuses mainly on increasing the codebook mapping performance. The codebook mapping performance is enhanced using predictive codebook mapping and optimal codebook interpolation. The predictive codebook mapping smoothes the high-band features over time, which helps in the reduction of perceptually noise artifacts. The optimal codebook interpolation improves the mapping performance.

In [10], the proposed ABE approach considers the spectral shifting excitation extension method, LPC features for the narrowband and wideband spectral envelopes, and VQ codebook approach. The spectral shifting method was implemented using two fixed modulation frequencies, 3.3 kHz and 4.7 kHz, with appropriate filtering to avoid overlapping. While in [12], a fixed modulation frequency is chosen for estimating the high-band excitation signal extension. Moreover, some additional narrowband information is taken as normalized short time energy and gradient index. Finally, predictions of the wideband features from the VQ codebook are enhanced by using a two-stage classification method. The normalized short time energy and the gradient index indicate voiced and unvoiced sounds in a better way. The two-stage classification method reduces artifacts in the synthesized wideband signal.

In [11], the high-band excitation is generated using the spectral translation method, which uses the fixed modulation frequency of 3.4 kHz. The narrowband information is taken as auto-correlation coefficients, zero-crossing rate, normalized frame energy, gradient index, kurtosis, and spectral centroid. Zero-crossing rate, kurtosis, and spectral centroid characteristics help in the better indication of the voiced and unvoiced sounds, plosive and vocal sounds, and fricative sounds, respectively. The high-band spectral information is represented by the cepstrum features and estimated by using the HMM with the GMMs model. The ABE approach proposed in [13] is almost similar to the ABE approach proposed in [11] except some modifications. The ABE approach is analyzed for the MFCC narrowband features apart from the auto-correlation coefficients. The MFCC features perform better than the auto-correlation coefficients. The modulation frequency has been chosen 8 kHz. The spectral floor suppression technique (SFS) is used to control the synthesized energy in the high-band. Also, it helps in the suppression of the noise artifacts synthesized in the estimated high-band speech signal. In [13], different statistical models have been analyzed wherein the DNN model performs well.

The ABE approach proposed in [14] uses the BP-MGN excitation extension method, the LSF features and pitch gain as the narrowband features, the LSF features and modulation gain as the high-band features, and separate GMM models for estimating the LSF features and modulation gain. The modulation gain is utilized to set the energy of the synthesized high-band signal.

In [15], the ABE approach is implemented using the full-wave rectification along with a spectral whitening filter for the excitation extension, cepstrum features for the narrowband and wideband spectral envelopes, and VQ codebook approach. The spectral whitening filter is used to obtain the flat spectrum of the excitation.

The bandwidth extension in [25] is performed by using the spectrum folding method for the excitation extension, the cepstrum features, normalized frame energy, and gradient index as the narrowband features, the LPC features as the wideband features, and the HMM with GMMs as a statistical model.

In [16], the proposed ABE framework uses the adaptive spectral double shifting technique with an excitation synthesis filter for obtaining the wideband excitation signal, the LSF features for the narrowband and wideband spectral envelopes, tilt filter, linear mapping matrix, and DNN model. It uses two successive LP analysis filters for obtaining the narrowband whitened excitation signal. The first LP analysis filter is applied to the narrowband speech for producing the narrowband excitation. The second is applied to the narrowband excitation for generating the narrowband whitened excitation signal. Further, the adaptive spectral double shifting technique is applied to the narrowband whitened excitation signal for obtaining the wideband whitened excitation signal. The wideband excitation signal is generated by passing the wideband whitened excitation signal through an excitation synthesis filter. The wideband excitation signal is fed to the tilt filter for reducing the over-energy artifacts. The excitation synthesis filter is estimated using the linear mapping matrix.

A few strategies for ABE are different from the source-filter model. The ABE method based on temporal envelope modeling is developed in [33]. In the temporal envelope modeling (TEM), the speech signal is decomposed into a temporal envelope and a fine structure. The temporal envelope represents the temporal energy contour. The fine structure represents rapid fluctuations. The high-band signal is estimated using the temporal envelope modeling. The high-band signal is derived by summing the sub-band signals for ABE. Each sub-band signal is obtained by multiplying the temporal envelope with the fine structure. The temporal envelope information of each sub-band is estimated using the GMM model, while the fine structure is directly estimated using the full-wave rectification method and narrowband signal. The temporal envelope modeling is used to achieve a better perceptual cue of the HB information. In [34], the ABE approach is proposed based on sparse representation of speech signals. It employs sparse coding over different dictionaries corresponding to voiced and unvoiced portions of the input speech. The ABE approach proposed in [35] is based on the amplitude modulation and frequency modulation (AM-FM) model. This model considers an AM-FM signal to represent each speech resonance. The speech signal is expressed as the sum of N (finite integer) successive AM-FM signals. A multi-band analysis scheme is used to isolate the AM-FM signals (resonance isolation) of the speech signal. It uses a bank of band-pass filters centered at each spectral peak (resonance) with an appropriate bandwidth for resonance isolation. The missing high-frequency bands (high-frequency AM-FM signals) are estimated using an iterative adaptation algorithm based on a least mean square error criterion.

Some ABE approaches directly estimate the high-band spectral information. In [36], the log-spectral power magnitude is taken to represent the high-band and narrowband information. At the same time, in [37], additional attributes such as MFCC, LSF, and band-pass voicing coefficient (BPVC) are used to capture narrowband information. Further, the high-band spectral magnitude information is estimated using the DNN model. The phase of the high-band spectrum is obtained by imaging the phase of the narrowband spectrum. In [38], the spectral magnitude is taken for representing the wideband and narrowband information. It uses a joint dictionary training approach for ABE. In joint dictionary training approach, dictionaries for the narrowband and wideband spectrograms are trained in a coupled manner, which capture the sparsity of the narrowband and wideband spectrograms using the same sparse coefficient. In [23], the constant Q-transform feature is used to represent narrowband and high-band information. The GMM model is used for predicting the high-band information. In [39], the log-spectral magnitude represents the narrowband information, while the cepstrum features represent the high-band spectral magnitude information. The phase for the high-band spectrum is obtained by shifting the phase of the narrowband spectrum. The DNN model is used for predicting high-band information.

1.3 Motivation, challenges, and our aims

The speech production model (SPM) is the most popular speech modeling used in speech coding, speech synthesis, speech recognition, speaker recognition, and speaker verification because of providing a more accurate estimation of speech parameters. Therefore, we consider this speech modeling scheme in our work. The spectral envelope is modeled using all-pole modeling in existing methods based on the speech production model [13, 16, 17]. According to the speech production theory, the spectral envelope (speech production filter) can be represented accurately by a pole-zero model (signal model[1]) [40]. However, all-pole modeling may not be sufficient to accurately represent envelopes of sounds like fricatives, nasals, laterals, and the burst interval of stop consonants due to the presence of zeros in the frequency response of the speech production filter [40]. In all-pole modeling, an invertible all-pole model (LP synthesis filter) represents the speech production filter, which can be obtained by the linear prediction (LP) method [5]. LP coefficients (LPC) representing denominator polynomial coefficients of the all-pole model can be taken directly for training the model in a machine learning technique and estimated directly. Also, the LP synthesis filter and the LP analysis filter are inverse to each other. Therefore, the all-pole modeling is simple. But, the LP coefficients are highly sensitive to the error obtained in their predictions [5]. Because error obtained in LP coefficients may produce an unstable all-pole model. This problem is tackled by transforming LP coefficients into LSF domain. In our work, we are going in a new direction where we obtain a stable synthesis filter with utilizing pole-zero modeling.

This thesis aims to utilize pole-zero modeling for representing the speech production filter. In pole-zero modeling, the analysis filter can not be directly obtained by inverting a causal and stable synthesis filter. In the inversion process, zeros of the synthesis filter will be poles of the analysis filter. Zeros lying outside of the unit circle of the synthesis filter makes the unstable analysis filter. Also, we have restricted to have equal number of poles and zeros in the synthesis filter. Therefore, we find a synthesis filter corresponding to an analysis filter by minimizing the error e shown in Figure 1.2. In Figure 1.2, an error system is made by combining the

[1] In this thesis, terms signal model and pole-zero model are used interchangeably.

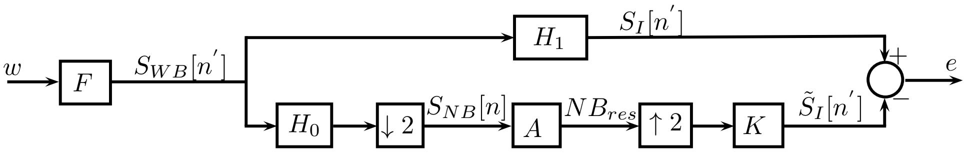

Figure 1.2: A general architecture of the error system.

transmitter set-up used for generating the narrowband signal, the bandwidth extension process used for estimating a signal of interest on the receiver side, and the generation process of the signal of interest. In Figure 1.2, the bottom part of the error system is a simplified model of a general ABE process. At the transmitter side, the wideband signal $S_{WB}[n']$ passes through a low-pass filter H_0 and downsampler $\downarrow 2$ (by a factor 2). The resulting signal $S_{NB}[n]$ is the transmitted narrowband signal. n and n' are the sample indexes for 8 kHz and 16 kHz sampled signals, respectively. The downsampler with the downsampling factor 2 is defined as

$$\psi[n] = y[2n], \quad n = 0, 1, 2, 3, 4, 5.... \tag{1.2}$$

where y and ψ are discrete input and output, respectively. At the receiver end, the received signal $S_{NB}[n]$ (assuming no loss) passes through the LP analysis filter A, upsampler (by a factor 2), and the (yet to be determined) synthesis filter K to reconstruct the signal of interest S_I. Three things are critical here. The first is the pole-zero model F of prior information. Specifically, it contains the spectral envelope information of the wideband speech signal. w is an input with know features (with finite energy, specifically $w \in l^2(\mathbb{Z}, \mathbb{R}^n)$. The second is the generation process of the signal of interest $S_I[n']$, which is represented by the system H_1. For example, we may like to focus only on reconstructing the high-band. In that case, we can take H_1 as a high pass filter. However, if one would not like to lose the wideband focus, we can take H_1 as the identity. There are many such situations. In this thesis, we experimented with three such signal interest conditions. The third thing is the design of unknown K. Solution of the error system can be obtained using the methods explained in the H^∞ sampled-data control theory [41–45]. The sampled-data system usually means a hybrid system that contains a combination of discrete-time (including multi rate systems) and continuous-time signals (see, e.g., [46]). In this thesis, we mainly consider multi-rate systems, which can be viewed as a special case of a sampled-data system in a broad sense. Also, we use methods explained in sampled-data control theory literature, especially the work of Nagahara and Yamamoto [47]. Therefore, we use the word sampled-data control often instead of any other term like discrete-time control or multi-rate system. We use the results given in [47–51] for obtaining a solution. For a quick summary of the sampled-data system theory, see Appendix A. The theory uses the inter-sample information optimally using the lifting technique [47]. This theory is also used for providing a robust solution in case of modeling uncertainties [52]. The robust solution provides some protection against uncertain and unknown speech signals in practical scenarios [3].

Define

$$G := \begin{bmatrix} H_1 \\ H_0 \end{bmatrix} F =: \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \tag{1.3}$$

We call G_1 and G_2 are the signal models (or pole-zero models). Depending upon the H_1 and H_0, we have different types of signal models G_1 and G_2. In this thesis, three types of H_1 are proposed. We have experimented with the three types of H_0 also. Moreover, to have more realistic scenarios, in the few experiments, we insert a speech codec block before the LP analysis filter A, as given in [13, 39]. This means adaptive multi rate (AMR) speech codec at 12.2 kbps. See [39, 53, 54] for details.

The speech signal's non-stationary nature does not allow the sampled-data system theory to be used as it is. This is because the theory works only for the linear time-invariant (LTI) system. It is well-known that the speech is stationary for the small duration of around 10-30 ms, and hence, we can obtain an LTI model. This means we will obtain a synthesis filter for the small duration, which is not a desirable situation computationally and practically. To this end, we propose to use regression to predict the synthesis filter. We have used machine learning (esp. GMM/DNN) for this task.

1.4 Contributions of the Thesis

Each contribution of the thesis is explained in subsequent chapters with sufficient details. Here, we list down the significant contributions. As mentioned in the previous section, the major contribution of this thesis is the *formulation of the ABE problem as an extension of the sampled-data system theory.* The formulation part requires a lot of innovation to perform as per the current standard. We now list the contributions provided in this thesis:

Chapter 2: A framework is developed for ABE using H^∞ sampled-data system theory on a simplified setup. This framework is the basis for the further chapters in this thesis. For simplification, we drop the anti-aliasing low pass filter before downsampling, i.e., $H_0 = 1$ in (1.3). Hence, the narrowband signal generated at the transmitter is no longer perfect; it includes aliasing distortion. It is to note that the aliased narrowband signals may have less intelligibility, but these are hypothesized to establish the better conditional dependence between narrowband and wideband information. The full wideband signal is estimated by using a synthesis/interpolation filter due to aliasing in the narrowband signal. Therefore, the proposed ABE approach considers the wideband signal modeling, i.e., $H_1 = 1$ in (1.3). A signal model is used to represent the wideband signal information. In this context, a novel error system is proposed by taking the aliased narrowband signal generation process, bandwidth extension process, and reference wideband signal. Further, solution of the novel error system is obtained using the methods explained in the H^∞ sampled-data system theory [41–45]. The solution is the synthesis filter, which is used in the bandwidth extension process to estimate the full wideband signal. A large number of synthesis filters are required to reconstruct the whole speech signal in a practical scenario due to the fact that the speech signal is non-stationary. This problem is solved by using the two statistical modeling approaches such as the Gaussian mixtures model and feed-forward DNN. A drawback of this approach is not compatible with the existing transmitter setup.

Chapter 3: A new framework has been proposed for ABE using H^∞ sampled-data system theory that works with the current technologies. We have followed the ITU-T standards commonly used by the literature for better comparison [39, 53, 54]. Specifically, we have work with the band-limited narrowband (approximately 300-3400 Hz) signal encoded at 12.2 kbps [54], i.e., H_0 is the low pass filter in Figure 1.2. The proposed ABE approach considers wideband signal modeling. In this context, a signal model (pole-zero model) is used to capture the spectral envelope information of the wideband (50-7000 Hz) signal. A novel error system is proposed and built up by considering the narrowband signal generation process, bandwidth extension process, and reference wideband signal generation process. This error system is designed for taking the pole-zero information of a signal into account. Further, solution of the error system is obtained by using the methods explained in the H^{∞} sampled-data system theory. The solution of the error system is a synthesis filter, which is used in the bandwidth extension process. The synthesis filter has the narrowband envelope information as well as the high-band envelope information, but the narrowband envelope information is not needed. Therefore, the narrowband information is suppressed in the synthesis filter. The energy of the estimated high-band signal is controlled by using a gain adjustment technique and a spectral floor suppression technique [13, 31, 55]. A large number of synthesis filters and corresponding gains are required to reconstruct the whole speech signal in a practical scenario due to the fact that the speech signal is non-stationary. This problem is solved by using a DNN model, which provides a kind of compact form representing the information of synthesis filters and gains. The proposed ABE approach extends the encoded narrowband signal for a realistic scenario [2, 56]. The standard transmitter set-up (as described in [39, 53, 54]) is followed in this work. The error system is also adapted according to the standards. Subjective and objective analyses are performed by considering the two datasets using the DNN model.

Chapter 4: The major change in this chapter is to consider this high pass filter in the error system for better optimization. We know that ABE aims to extend the bandwidth of the narrowband (NB) speech signal (up to 4 kHz) to 8 kHz. In this chapter, a new ABE approach is proposed based on high-band signal modeling, i.e., H_1 is the high pass filter in Figure 1.2. In this context, a signal model is used to represent better the high-band (4-8 kHz) information of a signal. A novel error system is proposed and made by considering the narrowband signal generation process, bandwidth extension process, and reference/true high-band signal generation process. Solution of the error system is obtained by using the methods explained in the H^∞ optimization (as in the earlier chapters). The solution of the error system is a synthesis filter for the given signal model. The obtained synthesis filter has the high-band (HB) spectral envelope information. The synthesis filter is used in the bandwidth extension process for synthesizing the high-band (4-8 kHz) signal. The discrete Fourier transform (DFT) concatenation is performed to add the narrowband signal sampled at 16 kHz and the estimated high-band signal sampled at 16 kHz for removing the leaked information from the synthesis filter and non-ideal low pass filter. Gain adjustment is performed on the estimated high-band signal to make its energy equal to the true high-band signal. Non-stationary characteristics of speech signals generate assorted variety in synthesis filters and corresponding gains. For this, a deep neural network (DNN) is trained to estimate the synthesis filter and gain by using only the narrowband information. We analyze the performance of the DNN model on two datasets. Objective and subjective analyses are carried out on these datasets.

Chapter 5: In this chapter, we extend the ABE approach proposed in Chapter 4. The proposed ABE approach is based on the mapped high-band signal modeling (shifting the high-band frequencies in the narrowband region) and H^∞ optimization. Further, an error system is proposed for minimizing error in the case of mapped high-band signal modeling. The error system is built up by combining the narrowband signal generation process, bandwidth extension process, and reference signal generation process. The reference signal is then the mapped high-band signal or band-pass shifted signal, which has the original high-frequency components shifted in the narrowband region. A gain factor corresponding to the synthesis filter is computed and used for adjusting the energy levels of the estimated high-frequency components. The spectral floor suppression technique with slight modification [13] is utilized for controlling the noise artifacts present in the estimated high-frequency components. Speech signals have time-varying characteristics. Therefore, several synthesis filters and corresponding gains are needed for constructing the whole speech signal. Hence, two different deep neural networks (DNNs) are designed for estimating the synthesis filter information and gain factor. We design separate DNN models for modeling the synthesis filter and the gain factor. In addition, the gain factor is computed and modeled in such a way that the gain factor reduces the performance loss obtained due to error in the predicted synthesis filter.

2

A new paradigm in artificial bandwidth extension

Contents

2.1 A Proposed set-up based on wideband modeling for artificial bandwidth extension of speech signals 19

2.2 Experimental analysis and results 28

2.3 Conclusion 36

In this chapter, we are doing preliminary experiments to check the utility of H^{∞} sampled-data system theory solution in speech signal enhancement. Many ABE techniques rely on the common theme of decomposing the narrowband and high-band information (see [13, 16, 31]). This is because only the high-band information is missing at the receiver. On account of the decomposition of narrowband and high-band information at the transmitter, two challenges arise for the effective ABE of the narrowband speech signal: (i) weaker conditional dependence between narrowband and wideband specifically for the unvoiced frames of speech and (ii) need to adjust of energy level between the estimated high-band and retained narrowband speech signals [31, 55]. In different unvoiced frames of speech, narrowband information is almost the same, while high-band details vary. Therefore, it isn't easy to estimate the respective high-band information for given narrowband information of the unvoiced frame. To tackle these challenges, a new ABE framework is proposed in this work. The proposed work differs from the existing works in three aspects. First, the narrowband signal generated at the transmitter is no longer perfect. It includes aliasing distortion due to dropping the low pass filter before downsampling. It is to note that the transmitted aliased narrowband signals may have less intelligibility, but these are hypothesized to establish the better conditional dependence between narrowband and wideband information. This is because high-band information is reflected in the narrowband region after downsampling, which yields more variations among the narrowband features for the unvoiced speech. Second, the interpolation filter of the speech signal is estimated by using the H^{∞} optimization/filtering, which is recommended in the literature (especially in control) to handle variations in system models (in our case, the pole-zero wideband signal model) [52]. Third, a large number of interpolation/synthesis filters are required to reconstruct the whole speech signal in a practical scenario due to the fact that the speech signal is non-stationary. This problem is solved by using the two statistical modeling approaches such as the Gaussian mixtures model and feed-forward DNN model.

This chapter discusses the new ABE framework along with a practical method to use H^{∞} sampled-data system theory for artificial bandwidth extension. It is evident from the discussion above that the proposed ABE approach is not suitable for the existing transmitter setup due to dropping of the low pass filter. This chapter is based upon these papers [57, 58].

The remaining part of the chapter is organized as follows: Section 2.1 has a detailed discussion about the proposed set-up for ABE. Section 2.2 consists of the experimental results and analysis using the GMM and DNN models for the proposed method. Also, this section has a comparison of the proposed method with the baselines. Section 2.3 concludes the proposed work.

2.1 A Proposed set-up based on wideband modeling for artificial bandwidth extension of speech signals

This section describes the proposed ABE approach for an aliased narrowband speech signal sampled at 8 kHz. A basic block diagram for ABE is shown in Figure 1.1. As it can be observed in Figure 1.1, a pre-trained model is needed in advance. The pre-trained model is designed using a database of narrowband features and high-band features. The pre-trained model for the proposed ABE framework is designed in the training block, as shown in Figure 2.1. The training block is elaborated in Section 2.1.1. After designing the pre-trained model, ABE process uses the pre-trained model at the receiver side, as shown in Figure 1.1. As evident from Figure 1.1, the ABE process consists of four main processes: high-band features estimation, NB features extraction, bandwidth extension process, and narrowband signal reconstruction process. These processes are used in the estimation of wideband (WB) signal. However, the proposed ABE approach does not use the narrowband signal reconstruction process. The extension block in Figure 2.1 consists of a description of the proposed ABE approach for estimating the wideband signal corresponding to the aliased narrowband signal. The extension block is explained in Section 2.1.2.

2.1.1 Training block

The training block consists of three sequential processes: windowing and framing process, features extraction process, and modeling process. The windowing and framing process is performed to produce stationary speech signals, as explained in Section 2.1.1.1. The features

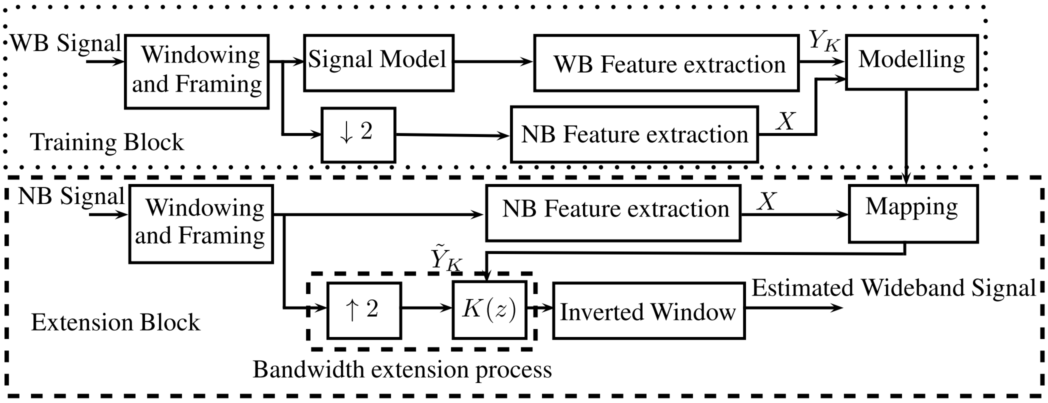

Figure 2.1: Block diagram consists of training of a model and extension of the narrowband signal.

extraction process computes two features: wideband feature vector $\mathbf{Y}_K$ and narrowband feature vector $\mathbf{X}$. These features are computed in Sections 2.1.1.2 and 2.1.1.4. These features are modeled using statistical models explained in Section 2.1.1.5.

2.1.1.1 Windowing and framing

It is a well-known fact that the characteristics of speech signals change with time (non-stationary) [3]. Hence, speech signals are segmented into frames, and these frames are considered as stationary signals. Here, speech signals are windowed into frames of 25 ms duration with 50% overlapping between adjoining frames using the Hamming window.

2.1.1.2 Wideband feature vector extraction

The wideband feature vector consists of the proposed synthesis/interpolation filter K information. Filter K is designed using the H^∞ optimization. For designing filter K, an error system is made by combining the wideband speech signal, narrowband generation process, and bandwidth extension process, as shown in Figure 2.2.

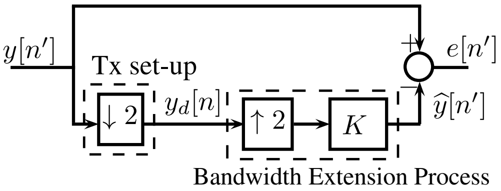

Figure 2.2: Error system set-up for reconstructing of a stationary speech signal.

In Figure 2.2, $y[n']$ represents the wideband stationary speech signal, $\downarrow 2$ is an ideal downsampler with the downsampling factor 2, $\uparrow 2$ is an ideal upsampler with the upsampling factor 2, $y_d[n]$ denotes the aliased narrowband stationary speech signal, and $e[n']$ denotes the error between the original wideband signal $y[n']$ and estimated wideband signal $\widehat{y}[n']$. n and n' denote 8 kHz and 16 kHz sample index, respectively. The signal $y[n']$ is downsampled by a factor of 2 at the transmitter (Tx) side to produce the narrowband signal $y_d[n]$. This narrowband signal generation process introduces distortion (aliasing) in the narrowband speech signal. Hence, our work is focused on estimation of the full wideband (0-8 kHz) signal at the receiver side. It means the signal of interest is the wideband signal. The bandwidth extension process is applied to the narrowband speech signal $y_d[n]$, which produces the estimated wideband signal $\widehat{y}[n']$. The synthesis filter K is designed in such a way that it minimizes the reconstruction error. We use system norm to measure the reconstruction error [59].

Now, we consider the signal modeling. The signal model represents the known characteristics of the signals. In this work, we assumed them linear discrete time-invariant (LDTI) systems. There are many ways to represent the LDTI system (see, e.g. [40]). In our work, pole-zero information about the original (to be processed) wideband signal $y[n']$ is extracted in form of a signal model F driven by external signal w_d. Then, a modified error system of Figure 2.2 is represented in Figure 2.3. In Figure 2.3, the signal $y[n']$ is an output of system F driven by an input signal $w_d[n']$ with known features (with finite energy, specifically $w_d \in \ell^2(\mathbb{Z}, \mathbb{R}^n)$). $y[n']$ can be voiced signal, unvoiced signal, or a combination of them. Note that, due to non-stationary nature of speech signal, there will be always modeling error in F (in the ABE process, we have adopted). To circumvent that problem, we use H^∞ norm and machine learning

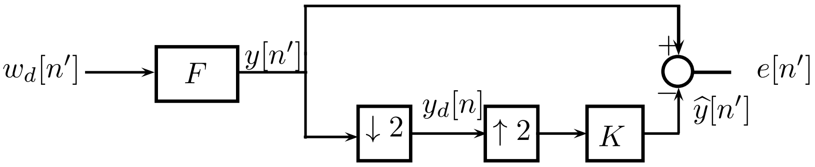

Figure 2.3: Proposed architecture of error system with considering signal modeling for reconstructing a stationary speech signal.

modeling techniques, which will be explained later. $F(z)$, which is the rational transfer function of F, is assumed to be a stable and causal transfer function. In Figure 2.3, F denotes both the signal models G_1 and G_2 defined in (1.3). Both the signal models G_1 and G_2 have the spectral envelope information of the wideband signal (16 kHz). When compared to Figure 1.2, we can easily see that $H_1 = 1$, $H_0 = 1$, and $A = 1$ in Figure 2.3. The signal of interest is the wideband signal $y[n']$ in the error system. Therefore, wideband modeling is used.

Performance index

The H^∞ system norm is used to minimize the reconstruction error. This is because this norm handles small modeling errors [52]. The H^∞-norm of a system $\mathcal{G}$ with input $\mathcal{X} \in l^2(\mathbb{Z}, \mathbb{R}^n)$ and output $\mathcal{Y} \in l^2(\mathbb{Z}, \mathbb{R}^m)$ is defined as (see, e.g., [1, 47, 52])

$$\|\mathcal{G}\|_\infty := \sup_{\mathcal{X}\neq 0} \frac{\|\mathcal{Y}\|_2}{\|\mathcal{X}\|_2}, \tag{2.1}$$

where $\|.\|_2$ represents the l^2-norm, and $\|.\|_\infty$ represents the H^∞-norm.

Problem formulation

To design optimal $K(z)$, the following optimization problem is solved.

Problem 1. *Given a stable and causal $F(z)$, design a stable and causal interpolation filter K_{opt} defined as*

$$K_{opt} := \arg\min_{K}(\|\mathbb{T}\|_\infty), \tag{2.2}$$

where $\mathbb{T} := F - K(\uparrow 2)(\downarrow 2)F$. $\mathbb{T}$ maps w_d to e (see Figure 2.3).

As mentioned earlier, the non-stationary behavior of speech signal introduces some uncertainty in estimation of the signal model $F(z)$ (in ABE process, we have adopted). In such a case, H^∞-norm optimization provides a robust solution against small modeling error in $F(z)$ [52].

Solution of Problem 1

This solution is essentially from [47, 48, 57, 60, 61]. Problem 1 is solved to design an optimal filter K_{opt}.

The error system $\mathbb{T}$ is converted into the generalized error system (see Figure B.1) as follows

$$\begin{aligned} G_1(z) &= F(z), \\ G_2(z) &= F(z), \\ G_3(z) &= 1, \\ K_d(z) &= K(z). \end{aligned} \tag{2.3}$$

Further, the solution of Problem 1 is obtained using the solution given for generalized error system in Appendix B.

Remark 1. *For downsampling by a factor N, see [47, 48].*

Filter K has an infinite impulse response (IIR). Practically, the IIR filter K can not be directly stored in a statistical model. Therefore, this filter is converted into an approximate finite impulse response (FIR) interpolation filter by truncating its Taylor series at the origin. The number of terms in the FIR interpolation filter is chosen 21 empirically, which is explained in Section 2.2.2. This FIR filter response is taken as the wideband feature vector $\mathbf{Y_K}$ in this work.

2.1.1.3 Computation of $F(z)$

The signal model $F(z)$ for a given stationary wideband signal is computed by the standard Prony's method based function available in MATLAB [62, 63]. The obtained model is causal and but may be unstable. To make it stable, those poles of the model, lying outside of the unit circle, are emulated inside by reciprocating their magnitudes without altering the phase [40]. Note that, the magnitude spectrum of $F(z)$ remains the same, however, the phase spectrum changes. This stabilizing process does not affect the perception of speech signals because the human auditory system is less sensitive to phase information [40]. Here, a question remains on choosing the number of poles and zeros in the signal model $F(z)$. Because poles and zeros in the signal model are not ideally the same for each speech frame [40]. Therefore, the number of poles and zeros is empirically calculated for each frame. In this context, we compute different H^∞ optimal synthesis filters $K(z)$ corresponding to different signal models $F(z)$ obtained by varying the number of poles and zeros (in the range of 11 to 81) in the MATLAB function. The number of poles is taken one more than the number of zeros. Further, each $K(z)$ is used in the estimation of the wideband signal. To this end, we calculate ℓ^2-norm of the error $e[n']$ for the given wideband signal $y[n']$ in Figure 2.3. Then, we choose the number of poles and zeros, which produces the minimum ℓ^2-norm of the error.

Remark 2. *For example, a speech signal of 100 ms duration is divided into frames of 25 ms duration with 50 % overlapping (see Section 2.1.1.1 windowing and framing). We obtain 7 frames. For each frame, we find an $F(z)$ by following procedure explained in section 2.1.1.3. Further, we get 7 $K(z)$ corresponding to 7 $F(z)$.*

2.1.1.4 Narrowband feature vector extraction

The narrowband information (narrowband features) is taken in four different ways, i.e., linear prediction coefficients (LPC) [64], line spectral frequencies (LSF) [65], linear frequency cepstral coefficients (Cepstrum) [13], and Mel frequency cepstral coefficients (MFCC) [55, 66]. These parameters are computed from the narrowband speech signal. The dimension of the narrowband feature vector is fixed to 10.

2.1.1.5 Modeling

This section has a description of statistical models. A statistical model is used to estimate the wideband feature vector $\mathbf{Y_K}$ using the narrowband feature vector $\mathbf{X}$. For this purpose, a pre-trained model is trained using the narrowband and wideband features. In our work, two types of statistical models are used, which are explained next.

Gaussian mixture model

A feature vector $\mathbf{Z} \in \mathbb{R}^{31}$ is formed by concatenating the narrowband feature vector $\mathbf{X}$ of dimension $\mathbb{R}^{10}$ and wideband feature vector $\mathbf{Y_K}$ of dimension $\mathbb{R}^{21}$. The feature vector $\mathbf{Z}$ is modeled by the Gaussian mixture model (GMM) to obtain the joint probability distribution function (pdf) of the narrowband feature vector $\mathbf{X}$ and wideband feature vector $\mathbf{Y_K}$ [67]. The pdf of $\mathbf{Z}$ is modeled by summing the weighted multivariate Gaussian distributions as follows

$$p(\mathbf{Z}|\lambda) = \sum_{i=1}^{M} w_i p(\mathbf{Z}|\mu_{\mathbf{z}_i}, \Sigma_{\mathbf{zz}_i}), \tag{2.4}$$

with w_i being the contribution of the i^{th} Gaussian distribution out of M clusters, and $p(\mathbf{Z}|\mu_{\mathbf{z}_i}, \Sigma_{\mathbf{zz}_i})$ denotes the corresponding Gaussian pdf. It is written as

$$p(\mathbf{Z}|\mu_{\mathbf{z}_i}, \Sigma_{\mathbf{zz}_i}) = \frac{1}{(2\pi)^{d/2}|\Sigma_{\mathbf{zz_i}}|^{1/2}} e^{-\frac{(\mathbf{Z}-\mu_{\mathbf{z_i}})^T \Sigma_{\mathbf{zz_i}}^{-1}(\mathbf{Z}-\mu_{\mathbf{z_i}})}{2}}, \tag{2.5}$$

with d dimension of feature vector $\mathbf{Z}$, and $\mu_{\mathbf{z}_i}$ and $\Sigma_{\mathbf{zz}_i}$ being the mean vector and covariance matrix of Gaussian pdf, respectively, and they are defined as

$$\mu_{\mathbf{z}_i} = \begin{bmatrix} \mu_{\mathbf{x}_i} \\ \mu_{\mathbf{y_k}_i} \end{bmatrix}, \tag{2.6}$$

$$\Sigma_{\mathbf{zz}_i} = \begin{bmatrix} \Sigma_{\mathbf{xx}_i} & \Sigma_{\mathbf{xy_k}_i} \\ \Sigma_{\mathbf{y_k x}_i} & \Sigma_{\mathbf{y_k y_k}_i} \end{bmatrix}, \tag{2.7}$$

where $\mu_{\mathbf{x}_i}$ and $\mu_{\mathbf{y_k}_i}$ are mean vectors of $\mathbf{X}$ and $\mathbf{Y_K}$, respectively. $\Sigma_{\mathbf{xx}_i}$ and $\Sigma_{\mathbf{y_k y_k}_i}$ are covariance matrices of $\mathbf{X}$ and $\mathbf{Y_K}$, respectively. $\Sigma_{\mathbf{xy_k}_i}$ and $\Sigma_{\mathbf{y_k x}_i}$ are cross-covariance matrices of $\mathbf{X}$ and $\mathbf{Y_K}$, respectively. For estimating the parameters of GMM model, Expectation-Maximization [67] algorithm is used, which gives the maximum likelihood solutions, i.e., maximize the probability of generating the feature vectors from the model. This leads to a joint pdf of $\mathbf{X}$ and $\mathbf{Y_K}$.

Further, the wideband feature vector is estimated using the joint pdf for the given narrowband feature vector $\mathbf{X}$. For this, a mapping function $f(\mathbf{X})$ is found by considering the minimum mean squared error (MMSE) criteria [68]. Mean squared error

$$\varepsilon_{mse} = E[\|\mathbf{Y_K} - f(\mathbf{X})\|^2], \tag{2.8}$$

is computed, where $\mathbf{Y_K}$ and $f(\mathbf{X})$ represent the original wideband feature vector and corresponding estimated wideband feature vector for the given narrowband feature vector $\mathbf{X}$, respectively. To solve (2.8), Bayesian estimation theory is used that gives a mapping function.

This mapping function is a conditional mean of $\tilde{\mathbf{Y}}_{\mathbf{K}}$ given $\mathbf{X}$ and defined as [69]

$$
\begin{aligned}
f(\mathbf{X}) =& E(\tilde{\mathbf{Y}}_{\mathbf{K}}|\mathbf{X}), && (2.9)\\
=& \sum_{i=1}^{M} \alpha_i(\mathbf{X})[\mu_{\mathbf{y_k}_i} + \Sigma_{\mathbf{y_k x}_i}\Sigma^{-1}_{\mathbf{xx}_i}(\mathbf{X} - \mu_{\mathbf{x}_i})], && (2.10)\\
\alpha_i(\mathbf{X}) =& \frac{w_i p(\mathbf{X}|\mu_{\mathbf{x}_i}, \Sigma_{\mathbf{xx}_i})}{\sum_{l=1}^{M} w_l p(\mathbf{X}|\mu_{\mathbf{x}_l}, \Sigma_{\mathbf{xx}_l})}. && (2.11)
\end{aligned}
$$

The weighting function $\alpha_i(\mathbf{X})$ is a posterior probability of i^{th} component in the Gaussian mixture distribution from which feature vector $\mathbf{X}$ is generated. $\tilde{\mathbf{Y}}_{\mathbf{K}}$ denotes the estimated wideband feature vector, which is used in the artificial bandwidth extension of speech signal.

Deep neural network

A deep neural network (DNN) is used to estimate the wideband feature vector $\tilde{\mathbf{Y}}_{\mathbf{K}}$ for a given narrowband feature vector $\mathbf{X}$ [70]. DNN model has several parameters, such as activation function, number of hidden layers, number of units in each hidden layer, learning rate, regularization, optimizer, loss function, and mini-batch size, which need to be objectively checked empirically to design an optimal DNN model. A DNN feed-forward topology architecture is made up of N number of layers, consisting of $N-1$ hidden layers and one output layer. The output of the i^{th} layer for sample index n is defined as

$$
\mathbf{h_n^i} = f_i(\mathbf{W^i}\mathbf{h_n^{i-1}} + \mathbf{b^i}), \qquad 1 \le i \le N, \tag{2.12}
$$

where $\mathbf{W^i}$ and $\mathbf{b^i}$ signify the weight and bias parameters, respectively. $f_i(.)$ is a non-linear activation function, and $\mathbf{h_n^i}$ is the output of i^{th} layer. The output $(\mathbf{h_n^N})$ of the N^{th} layer yields the estimated wideband feature vector, and the input $(\mathbf{h_n^0})$ to the first layer is the narrowband feature vector. In (2.12), $\mathbf{W^i}$ and $\mathbf{b^i}$ are unknown parameters, which are initialized with some random value. Further, the mean squared error is considered as a loss function (α), which is minimized to obtain the optimal weight $\mathbf{W^i_{opt}}$ and bias $\mathbf{b^i_{opt}}$ values of each layer as described

$$\alpha = \frac{1}{T}\sum_{n=1}^{T} \|\mathbf{h^N_n} - \mathbf{Y^n_K}\|_2^2, \tag{2.13}$$

$$(\mathbf{W^i_{opt}}, \mathbf{b^i_{opt}}) = \underset{\mathbf{W^i},\mathbf{b^i}}{\arg\min}(\alpha), \tag{2.14}$$

with T being the mini-batch size and $\mathbf{Y^n_K}$ denotes the original wideband feature vector.

2.1.2 Extension block

In the extension block, the pre-trained models designed in Section 2.1.1.5 are used for the artificial bandwidth extension of the narrowband signal. The wideband signal is reconstructed using the three processes: windowing and framing, the same as done in Section 2.1.1.1, mapping process for estimating the wideband feature vector explained in Section 2.1.2.1, and estimation of wideband signal explained in Section 2.1.2.2.

2.1.2.1 Wideband feature vector estimation

The only narrowband signal is available at the receiver side. Therefore, the narrowband feature vector $\mathbf{X}$ is computed using the narrowband signal, as done in Section 2.1.1.4. Then, X is fed to the pre-trained model, which maps the feature vector X into the estimated wideband feature vector $\mathbf{\tilde{Y}_K}$. Mapping process uses (2.11) and (2.12) in the cases of the GMM model and DNN model, respectively.

2.1.2.2 Wideband signal estimation

The estimated wideband feature vector $\mathbf{\tilde{Y}_K}$ has the filter K information used in the bandwidth extension process. The narrowband signal is upsampled by a factor of 2 and then passed through the interpolation filter $K(z)$. Subsequently, the resulting signal is multiplied by the reciprocal of the Hamming window to estimate the wideband speech signal. Further, the overlapped portion of two adjacent frames is estimated by averaging the overlapped parts of the estimated wideband stationary signals. In other words, the weighted overlap-add method (WOLA) is applied to reconstruct full speech signal [71, 72].

2.2 Experimental analysis and results

In this section, experiments are conducted to establish the correctness and effectiveness of the proposed approach. Section 2.2.1 has a description of speech datasets used for evaluating the proposed approach. Section 2.2.2 has an objective analysis, which is done for evaluating the proposed approach. In objective analysis, we analyze the performance of the proposed approach to know the effectiveness of the proposed filter K. Besides, experiments are performed to decide the dimension of the wideband feature vector, show the proposed approach performance using two types of pre-trained models, and compare the proposed approach with two baselines. Section 2.2.3 consists of the subjective evaluation of extended speech files.

2.2.1 Databases

The proposed approach is evaluated on the TIMIT [73] and RSR15 [74] datasets. Both the datasets contain the recorded speech files at a sampling rate of 16 kHz. The TIMIT dataset is already segmented into train and test sets. The train set is used to train the model, while the test set is considered as a validation set. A new test set is made by taking speech files from the RSR15 dataset. This new test set has the speech files uttered by 4 female and 3 male speakers. The consideration of test set from a different database leads to more generalized results.

2.2.2 Objective analysis

In this work, several standard objective speech quality measures such as mean square error (MSE) [75], signal to distortion ratio (SDR) [76], log likelihood ratio (LLR) [3, 77], upper-band (4-8 kHz) logarithmic spectral distance (LSD$_{UB}$), full-band (0-8 kHz) logarithmic spectral distance (LSD$_{FB}$) [39, 78], narrowband MOS-LQO (mean opinion score listening quality objective) [79, 80], and wideband MOS-LQO [81, 82] are chosen for examining the quality of artificially extended speech signals. The mathematical formulations of these measures are given in Appendix C.

Further, we analyze the objective measures of extended speech signals obtained using the output signals at various parts of the bandwidth extension approach. Outputs of the upsampler and IIR interpolation filter are separately used to estimate wideband signals. This process is conducted by using the oracle filter K directly in the extension block of Figure 2.1. It means the narrowband speech signal is enhanced by applying the oracle IIR interpolation filter K on the upsampled narrowband signal. For this analysis, we take some speech files from the validation set. The objective measures are listed in Table 2.1 for extended speech signals estimated using the upsampler and oracle IIR interpolation filter K. Here, the interpolation filter K improves all the objective measures significantly.

Table 2.1: Performance comparison of extended speech signals enhanced by applying the upsampler (without applying filter K) and the oracle IIR interpolation filter K in Figure 2.1 on the speech files taken from the validation set.

Output subblock	MSE ($\times 10^{-5}$)	SDR	LLR	NB MOS-LQO	LSD$_{FB}$	LSD$_{UB}$	WB MOS-LQO
Upsampler	81.1673	3.01	1.4254	3.5044	11.3135	14.0124	1.0666
Interpolation filter K	4.8634	15.81	0.6547	3.8047	7.6220	9.1764	2.0155

Moreover, filter K has an infinite impulse response. It is transformed into an approximate FIR filter by truncating the Taylor series. For deciding the length of the FIR filter, objective measures are computed for enhanced speech files, which are enhanced by using the FIR filters of different lengths. In Table 2.2, the objective measures improve with increasing the number of terms present in the FIR filter, but slowly after the length 21. Hence, the filter length is set to 21.

Table 2.2: Performance evaluation for some speech files taken from the validation set in condition of direct implanting FIR filter K (oracle K) in Figure 2.1 for ABE.

Number of terms	MSE ($\times 10^{-5}$)	SDR	LLR	NB MOS-LQO	LSD$_{FB}$	LSD$_{UB}$	WB MOS-LQO
11	8.9405	13.18	0.7925	3.7450	8.2260	10.0941	1.6435
15	7.4762	13.74	0.7851	3.7521	8.1389	9.9537	1.7042
21	6.0912	14.79	0.7233	3.7782	7.9339	9.6452	1.8480
25	5.8136	15.06	0.7065	3.7810	7.8678	9.5378	1.8870
31	5.6043	15.25	0.6937	3.7854	7.8078	9.4545	1.9155

Further, we analyze the performances using the GMM model and DNN model in the proposed artificial bandwidth extension approach.

2.2.2.1 Performance evaluation using Gaussian mixture model

The GMM based regression technique is used to estimate an interpolation filter (wideband feature vector) for a given narrowband feature vector in the proposed approach. The GMM model with 128 mixtures is trained using the narrowband feature vectors and proposed wideband feature vectors. This experiment is performed for four types of narrowband attributes: LSF, LPC, Cepstrum, and MFCC. The proposed approach using the GMM model is tested on the test set. For this, objective measures for artificially extended speech files belonging to the test set are computed for the narrowband attributes, as listed in Table 2.3.

Table 2.3: Performance evaluation by using 128 GMMs on the test set.

Features	MSE ($\times 10^{-4}$)	SDR	LLR	Narrowband MOS-LQO	LSD$_{FB}$	LSD$_{UB}$	Wideband MOS-LQO
LSF+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	**3.4667**	**11.17**	**0.6063**	**3.5653**	**7.9945**	**9.9930**	**2.1970**
LPC+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	3.6206	10.73	0.6722	3.5629	8.4141	10.5752	2.0598
Cepstrum+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	3.4719	10.86	0.7192	3.5524	8.7476	10.9760	2.0211
MFCC+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	3.6033	10.90	0.6385	3.5642	8.2438	10.3218	2.1362

The objective measures are analyzed for all the narrowband features. LSF narrowband features produce the best performance in comparison to the other narrowband features.

2.2.2.2 Performance evaluation using deep neural network

DNN topology is used to estimate interpolation filter coefficients. Some preliminary experiments are done to decide the parameter values for DNN topology with fixing the narrowband features. An optimal DNN architecture is designed by optimizing its parameters over the fixed LSF narrowband features. AdaMax (adaptive moment estimation based on the infinity norm) [83] optimizer is used to update the weights of network by applying L_2 regularization empirically [70]. Experimentally hyper-parameters such as mini-batch size, epochs, learning rate α, decay rates β_1 for the first-moment estimate, and β_2 for the second-moment estimate over a broad range are set to 200, 50, 0.01, 0.9, and 0.999, respectively. Mean and variance normalization (MVN) is applied to the features by using the statistics obtained for the training set. Also, batch normalization before activation function is applied to each hidden layer. The

Table 2.4: Performance evaluation on the validation set for different DNN topologies by varying the number of hidden layers (N_{HL}) and the number of units (N_U), and ReLU activation function in hidden layers, linear activation function in the output layers, LSF narrowband features and *AdaMax* optimizer.

Topology with ReLU activation functions		Performance on validation set						
N_{HL}	N_U	MSE ($\times 10^{-5}$)	SDR	LLR	NB MOS-LQO	LSD_{FB}	LSD_{UB}	WB MOS-LQO
2	512	3.3349	15.19	0.7082	3.6948	7.7229	9.3855	1.8985
2	1024	3.3347	15.20	0.7074	3.6964	7.7202	9.3838	1.8998
3	128	3.3376	15.19	0.7053	3.6935	7.7166	9.3749	1.9007
3	256	3.3386	15.20	0.7046	3.6966	7.7131	9.3714	1.9024
3	512	3.3453	15.19	0.7055	3.6963	7.7162	9.3761	1.9012
3	1024	3.3521	15.19	0.7064	3.6981	7.7207	9.3814	1.8996
4	128	3.3292	15.21	0.7033	3.6908	7.7113	9.3678	1.9043
4	**256**	3.3174	15.23	0.7023	3.6916	7.7084	9.3633	1.9081
4	512	3.3247	15.22	0.7025	3.6928	7.7097	9.3653	1.9073
4	1024	3.3411	15.20	0.7042	3.6924	7.7175	9.3768	1.9038

ReLU activation function is used in hidden layers, and the linear activation function is used in the output layer. Performances of different DNN topologies on the validation set are tabulated in Table 2.4. Overall good performance on the validation set is acquired by four hidden layers and 256 hidden units. Next, this architecture is trained by changing mini-batch sizes, and performance on the validation set is tabulated in Table 2.5. It is observed that performance using the mini-batch size 50 is obtained better than other mini-batch sizes. We use mini-batch size 50 in further experiments. The optimal DNN architecture of 4 N_{HL} and 256 N_U is fixed

Table 2.5: Performance evaluation on the validation set for the DNN model designed using 4 hidden layers and 256 units in each hidden layer for different batch sizes.

Mini-batch Size	MSE ($\times 10^{-5}$)	SDR	LLR	NB MOS-LQO	LSD_{FB}	LSD_{UB}	WB MOS-LQO
200	3.3174	15.2280	0.7023	3.6916	7.7084	9.3633	1.9081
150	3.3244	15.2186	0.7022	3.6906	7.7076	9.3628	1.9070
100	3.3197	15.2241	0.7023	3.6909	7.7064	9.3639	1.9054
50	3.3170	15.2285	0.7020	3.6901	7.7048	9.3630	1.9048
40	3.3304	15.2067	0.7033	3.6908	7.7061	9.3720	1.9035

Table 2.6: Performance evaluation on the test set by the proposed approach using the DNN model designed using 4 hidden layers and 256 units in each hidden layer for different narrowband features.

Features	MSE $(\times 10^{-4})$	SDR	LLR	NB MOS-LQO	LSD$_{FB}$	LSD$_{UB}$	WB MOS-LQO
LSF+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	3.2783	11.61	**0.6350**	3.5643	**8.1894**	**10.2186**	**2.2048**
LPC+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	**3.2677**	**11.62**	0.6487	**3.5660**	8.2687	10.3268	2.1837
Cepstrum+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	4.2454	9.75	0.9356	3.4481	9.6169	11.8401	1.7943
MFCC+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	3.5402	11.20	0.6525	3.5579	8.2966	10.3693	2.1337

for all further experiments.

Moreover, the optimal DNN architecture is trained for the other narrowband features. The proposed approach using the DNN model is tested on the test set. It is done by computing the objective measures for the artificially extended speech files belonging to the test set. The objective measures are listed in Table 2.6 for the narrowband features. It is observed that LPC narrowband features yield better MSE, SDR, and narrowband MOS-LQO than the other narrowband features. The rest of the objective measures in the majority of the cases are obtained better for LSF narrowband features.

Furthermore, the objective measures are analyzed for the voiced speech and unvoiced speech of the test set separately. For this, speech signals are segregated into two fundamental parts: voiced speech and unvoiced speech by a glottal activity detection (GAD) method [84, 85]. The performance is analyzed for the voiced speech and unvoiced speech separately. Table 2.7 and Table 2.8 have the objective measures computed for the voiced speech and unvoiced speech taken from the test set, respectively, with varying narrowband feature definitions.

Table 2.7: Performance evaluation for the voiced speech extracted from speech files belonging to the test set using the DNN model designed using 4 hidden layers and 256 units in each hidden layer for different narrowband features.

Features	MSE $(\times 10^{-4})$	SDR	LLR	NB MOS-LQO	LSD$_{FB}$	LSD$_{UB}$
LSF+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	4.0418	**13.63**	**0.8924**	4.1548	**7.6249**	**9.7540**
LPC+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	**4.0240**	13.53	0.8988	**4.1556**	7.6530	9.8017
Cepstrum+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	7.4300	10.03	1.1916	4.0265	8.9441	11.3735
MFCC+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	4.4334	12.98	0.9238	4.1508	7.7619	9.9562

Table 2.8: Performance evaluation for the unvoiced speech extracted from speech files belonging to the test set using the DNN model designed using 4 hidden layers and 256 units in each hidden layer for different narrowband features

Features	MSE ($\times 10^{-4}$)	SDR	LLR	NB MOS-LQO	LSD$_{FB}$	LSD$_{UB}$
LSF+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	5.0273	**9.42**	**0.6365**	3.8445	**8.0101**	**9.6812**
LPC+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	5.0134	9.23	0.6473	**3.8451**	8.0846	9.7820
Cepstrum+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	**4.4418**	6.02	0.9379	3.7447	9.3435	11.1161
MFCC+$\tilde{\mathbf{Y}}_{\mathbf{K}}$	5.3710	8.80	0.6604	3.8406	8.1386	9.8574

The LSF narrowband feature yields the best SDR, LLR, LSD$_{UB}$ and LSD$_{FB}$ for the voiced speech and unvoiced speech. The LPC narrowband feature yields the best MOS-LQO for both the speeches. The MSE is obtained the better using the LPC narrowband feature for voiced speech and Cepstrum narrowband feature for the unvoiced speech. The LSF narrowband feature, among all the narrowband features, yields the best performance in the majority of the cases for voiced speech and unvoiced speech.

2.2.2.3 Performance comparison

The proposed method is compared with the two baselines: spectral translation [7,13,55] and cepstral domain approach [39]. Experimental conditions are kept the same as datasets, dimensions of wideband features, windowing, and DNN model. LSF features are used to represent the narrowband features and wideband features in the spectral translation technique. Also, this technique uses a gain factor calculated by following [55]. The cepstral domain approach uses the narrowband magnitude spectrum as the narrowband feature and cepstral coefficients as the wideband feature [39].

Moreover, these techniques are implemented by using the low pass filter for generating the narrowband signal, i.e., $H_0 \neq 1$. Here, the low pass filter is a non-causal FIR filter defined in [2]. Cut off frequency of the LPF filter is 3660 Hz. The length of this filter is 118. The non-causality of this filter introduces a delay in transmission. Objective measures are listed in Table 2.9 for the proposed approach and baselines implemented using the same DNN model.

As seen in Table 2.9, the baselines improve LLR, NB MOS-LQO, LSD$_{FB}$, LSD$_{UB}$, and the

Table 2.9: A comparison of the objective measures computed on the test set speech files for the proposed approach and the baselines

Methods	MSE ($\times 10^{-4}$)	SDR	LLR	NB MOS-LQO	LSD$_{FB}$	LSD$_{UB}$	WB MOS-LQO
Pure narrowband signal	9.6999	5.1491	1.2166	4.1786	12.6584	16.1126	**2.5179**
Spectral translation	9.9900	4.99	0.7945	4.3457	9.4882	10.9215	2.3762
Cepstral domain	9.7986	5.09	0.7336	**4.4058**	9.4486	11.1373	2.4446
Aliased narrowband signal	10.1760	4.9720	1.3540	3.4529	12.7022	16.0640	2.1011
Proposed method	**3.2783**	**11.61**	**0.6350**	3.5643	**8.1894**	**10.2186**	2.2048

proposed approach improves MSE, SDR, LLR, NB MOS-LQO, LSD$_{FB}$, LSD$_{UB}$, and WB MOS-LQO when compared with their respective narrowband signals. Also, the proposed method improves all the objective measures except the narrowband and wideband MOS-LQO values when compared to the baselines. The NB and WB MOS-LQO values are obtained better by the existing methods. It may be due to the available original narrowband information. In the baselines, the narrowband signal is generated by using the low pass filter. Therefore, the narrowband information does not alter. As a result, NB and WB MOS-LQO values are obtained better by the baselines than the proposed method.

Moreover, spectrogram of a speech file taken from the test set is analyzed. Figure 2.4 **(a)**, **(b)**, **(c)**, and **(d)** illustrate spectrogram of the reference speech signal, extended speech signals by the proposed approach, spectral translation technique, and cepstral domain approach, respectively. As observed in Figure 2.4, the spectrogram of the extended speech signal has more difference around 4 kHz from the original spectrogram for the baselines than the proposed method. It has happened because of the energy levels adjustment issue around 4 kHz in the existing methods. It is observed around 0.9 secs and 0.77 secs in Figure 2.4 that the estimated high-band information is more close to the original high-band information by the proposed method than the baselines. However, the estimated high-band information around 7-8 kHz and during 0.40-0.55 secs in Figure 2.4 is observed more than the original information by the proposed method when compared with the baselines.

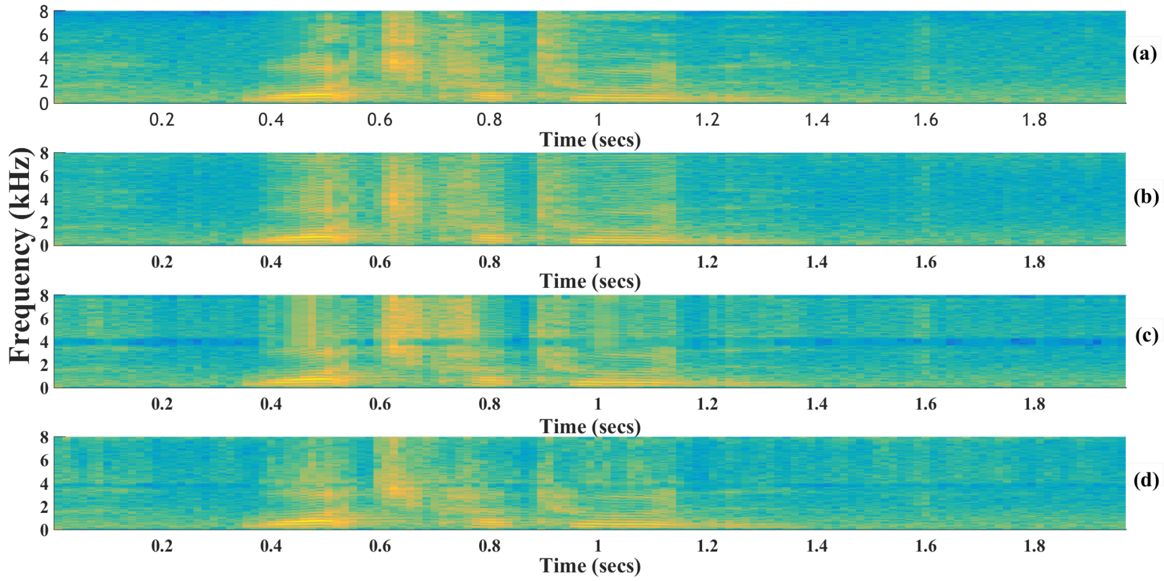

Figure 2.4: Spectrogram of **(a)** Original wideband signal, **(b)**, **(c)**, and **(d)** reconstructed wideband signal by the proposed method, spectral translation, cepstral domain, respectively.

2.2.3 Subjective listening test

Subjective assessment is done according to the ITU-T P.800 [86, Annex E] for examining the speech quality. This task is conducted for the extended speech signals obtained by the proposed method, spectrum translation technique, and cepstral domain approach using the same DNN architecture. Extended speech files by the proposed method are rated with respect to extended speech files by the existing methods. Ten pairs of extended speech signals belonging to the test set are randomly chosen for these methods, i.e., 60 files total. Then, twelve listeners were asked to give a mean opinion score (MOS) value between -3 (much worse) to 3 (much better). The ages of these listeners are between 23 to 32 years. These listeners do not have any hearing impairment and understand well English language. They were permitted to listen the speech files more than once. Further, 95% confidence interval (CI) is computed for measuring statistical significance. Then, the comparison mean opinion score (CMOS) and 95% confidence interval (CI) are listed in Table 2.10. Our proposed method improves CMOS significantly by 0.9375 and 1.5875 points in comparison to the spectral translation technique and cepstral

Table 2.10: Subjective assessment on artificially extended speech files belonging to the test set by the proposed method with respect to the baselines

Conditions	CMOS	CI
Spectral translation vs. Proposed method	0.9375	[0.7569 1.1181]
Cepstral domain vs. Proposed method	1.5875	[1.3630 1.8120]

domain approach, respectively. Unvoiced phonemes are perceived better in the extended speech files using the proposed method than the baselines.

2.3 Conclusion

A new framework (which capitalizes on artificially introduced non-ideality in the narrowband signal) is proposed for the artificial bandwidth extension of speech signals. In our proposed framework, the transmitter set-up is different from the existing transmitter set-up, which helps mainly in identifying the high-frequency components for the unvoiced speech. The discrete interpolation filter is obtained by using a signal model with the help of H^∞ optimization. The obtained rational stable and causal interpolation filter is converted into an FIR filter empirically. This FIR filter is taken as the wideband feature. Experiments are performed by considering four types of narrowband features: LSF, LPC, MFCC, and Cepstrum. Estimation of wideband feature for a given narrowband feature is conducted by two different machine learning modeling techniques: GMM and DNN. Performance is analyzed on the test set speech files taken from the RSR15 database by computing the standard objective measures: SDR, MSE, narrowband MOS-LQO, LLR, LSD$_{FB}$, LSD$_{UB}$, wideband MOS-LQO, and subjective listening test. Also, the objective measures are analyzed for the voiced speech and unvoiced speech separately. The proposed approach obtains the better LSD$_{UB}$ and LLR for the unvoiced speech than the voiced speech. The proposed approach improves the objective measures except narrowband and wideband MOS-LQO values in comparison to the baselines using the DNN model. In the listening test, CMOS is achieved higher by the proposed method than the baselines.

3

Artificial bandwidth extension technique based on the wideband modeling

Contents

3.1 A proposed set-up based on wideband modeling for artificial bandwidth extension of speech signals 39

3.2 Experimental set-up and results 48

3.3 Conclusion 54

In the previous chapter, the ABE framework was not suitable for the existing technologies. Hence, this chapter proposes a new ABE approach using H^∞ sampled-data system theory that works with the current technologies. We have followed the ITU-T standards commonly used by the literature for better comparison [39, 53, 54]. Specifically, we have worked with the band-limited narrowband (approximately 300-3400 Hz) signal encoded at 12.2 kbps (see complimentary paper [87] without encoding) [54]. The proposed ABE approach considers wideband signal modeling. In this context, a signal model (pole-zero model) is used to capture the spectral envelope information of the wideband (50-7000 Hz) signal. A novel error system is proposed and built up by considering the narrowband signal generation process, bandwidth extension process, and reference wideband signal generation process. This error system is designed for taking the pole-zero information of a signal into account. Solution of the error system can be obtained using the methods explained in the H^∞ sampled-data system theory [41–45]. The solution of the error system is a synthesis filter, which is used in the bandwidth extension process. The synthesis filter has the narrowband envelope information as well as the high-band envelope information, but the narrowband envelope information is not needed. Therefore, the narrowband information is suppressed in the synthesis filter. The energy of the estimated high-band signal is controlled by using a gain adjustment technique and a spectral floor suppression technique [13, 31, 55]. A large number of synthesis filters and corresponding gains are required to reconstruct the whole speech signal in a practical scenario due to the fact that the speech signal is non-stationary. This problem is solved by using a DNN model, which provides a kind of compact form representing the information of synthesis filters and gains. The proposed ABE approach extends the encoded narrowband signal for a realistic scenario [2, 56]. The standard transmitter (as described in [39, 53, 54]) set-up is followed in this work. The error system is also adapted according to the standards. Subjective and objective analyses are performed by considering the two datasets using the DNN model. This chapter is based upon the paper [88], which is a modified version of results without encoding [87].

The rest of the chapter is organized as follows: Section 3.1 has the proposed approach used to enhance the narrowband signal. The proposed approach for ABE includes the designing of the pre-trained model and then the wideband signal estimation. Designing the pre-trained model involves the pre-processing of speech signals, features extraction, and training of the DNN model. The wideband signal estimation process consists of narrowband signal reconstruction, high-band feature vector and gain factor estimation, and high-band signal estimation. Section 3.2 has the experimental results and analysis using the proposed bandwidth extension approach. Also, the proposed approach is compared with the two baselines. Section 3.3 concludes the proposed method.

3.1 A proposed set-up based on wideband modeling for artificial bandwidth extension of speech signals

A basic block diagram for ABE is shown in Figure 1.1. It can be observed in Figure 1.1, a pre-trained model is needed in advance. Its designing process is explained in Section 3.1.1 for the proposed ABE approach. The ABE process uses the pre-trained model at the receiver side, as shown in Figure 1.1. As evident from Figure 1.1, the ABE process consists of four main processes: estimation of the high-band features, NB features extraction, bandwidth extension process, and narrowband signal reconstruction process. These processes play an important role in the estimation of the wideband signal. Furthermore, the proposed ABE approach uses an additional process to adjust the energy level of the estimated high-band signal. In this chapter, the (encoded) narrowband signal is enhanced by the proposed bandwidth extension approach. Section 3.1.2 has a description of the proposed ABE approach for estimating the wideband signal corresponding to the encoded narrowband signal.

3.1.1 Designing of the pre-trained model

Here, a DNN model is trained to design the pre-trained model, as depicted in Figure 3.1. The training process of the DNN model involves two main sequential processes, viz, features extraction and DNN model training by using extracted features, as shown in Figure 3.1 [70]. The features extraction process derives three attributes viz. high-band feature vector $\mathbf{Y_K}$,

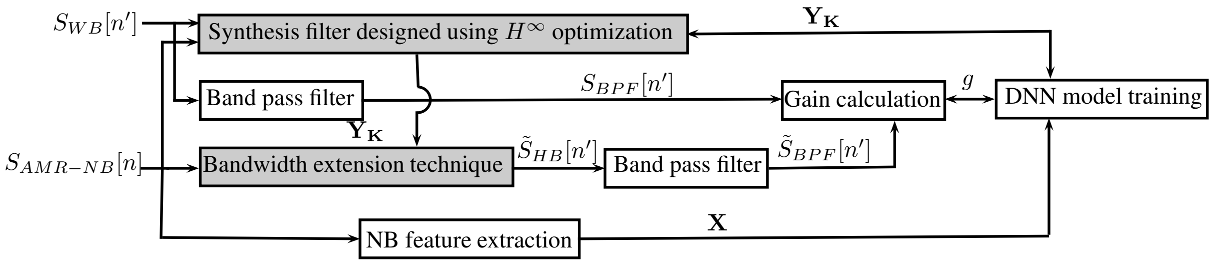

Figure 3.1: Block diagram Illustrating the training of the DNN model.

gain factor g, and narrowband feature vector $\mathbf{X}$. For computing these features, two input signals, wideband signal $S_{WB}[n']$ and encoded narrowband signal $S_{AMR-NB}[n]$ are needed in advance. Hence, these signals can be obtained using the processes described in Section 3.1.1.1. A description for computing the features $\mathbf{Y_K}$, g, $\mathbf{X}$, and training the DNN model is given in Section 3.1.1.2, Section 3.1.1.3, Section 3.1.1.4, and Section 3.1.1.5, respectively.

3.1.1.1 Pre-processing of speech signals

This section explains the process for producing the encoded narrowband signal $S_{AMR-NB}[n]$ at the transmitter side for realistic mobile telephone speech [2, 56]. Speech files sampled at 16 kHz are processed to produce the narrowband speech encoded at 12.2 kbps in narrowband telephonic communication [56, 89]. A process is drawn in Figure 3.2 for obtaining the encoded narrowband speech signal. In Figure 3.2, the original speech signal is filtered by the standard

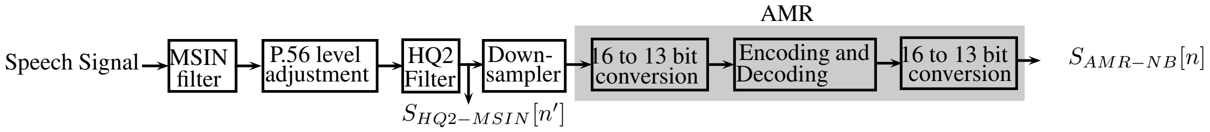

Figure 3.2: AMR coded narrowband signal generation process.

mobile station input (MSIN) high-pass filter [2] and subsequently scaled to an active speech level of -26 dBov [90]. The resulting signal is passed through another standard high-quality low pass filter (HQ2) [2] and then downsampled by a factor of 2. Thus, an obtained narrowband signal ($S_{NB}[n]$) is subjected to 16 to 13 bit conversion, encoding using the adaptive multi rate (AMR) narrowband speech codec at 12.2 kbps and subsequently decoding [89], and again

16 to 13 bit conversion, which gives the AMR coded narrowband signal $S_{AMR-NB}[n]$. This narrowband signal ($S_{AMR-NB}[n]$) is processed by the proposed ABE framework for synthesizing the frequency components up to 7 kHz.

The wideband signal $S_{WB}[n']$ is obtained by following Figure 3.3 wherein $S_{WB}[n']$ is generated by the standard P.341 filtering [2] of the original speech file sampled at 16 kHz and subsequently scaled to an active speech level of -26 dBov. The signal $S_{WB}[n']$ is taken as a reference signal for obtaining the synthesis filter and for performance evaluation. Here, n represent the sample index for 8 kHz sampled signal. n' represent the sample index for 16 kHz sampled signal. All these operations are performed for each 20 ms frame duration.

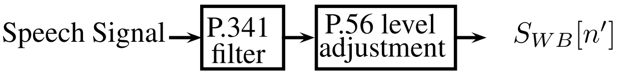

Figure 3.3: Wideband signal generation process.

Furthermore, each frame is multiplied by the Hanning window's square root with 50% overlap for adjacent frames for bandwidth extension. The whole speech signal is reconstructed by multiplying each estimated wideband signal (wideband frame) with the square root of the Hanning window and then combining resultant frames using the overlap-add method [71, 72].

3.1.1.2 High-band feature extraction

The high-band feature vector $\mathbf{Y_K}$ contains information of the proposed synthesis filter used in the proposed bandwidth extension process. The synthesis filter is designed by using the H^{∞} optimization. For this, an error system (Figure 3.4) is proposed by considering the narrowband signal generation process, bandwidth extension process used at the receiver side, and reference wideband signal $S_{WB}[n']$ generation process.

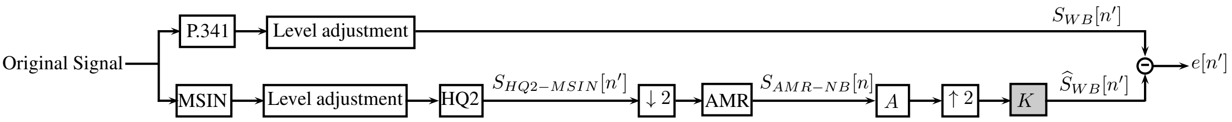

Figure 3.4: A proposed error system for wideband signal reconstruction.

In Figure 3.4, the signal $S_{HQ2-MSIN}[n']$ is obtained by passing the original signal through the

MSIN filter followed by -26 dBov level adjustment and the HQ2 low pass filter. ↓ 2 depicts a downsampler with a downsampling factor of 2. Analysis filter A (Figure 3.4) is the reciprocal of an all-pole model (order 16) of signal $S_{AMR-NB}[n]$ obtained by linear prediction (LP) analysis [5]. An output of filter A is the narrowband residual signal. AMR block (Figure 3.4) performs 16 to 13 bit conversion, encoding and decoding, and again 16 to 13 bit conversion operations. $\widehat{S}_{WB}[n']$ represents the estimated wideband signal. The error between the estimated and reference wideband signal is represented by $e[n']$. A filter K is obtained in such a way that it minimizes the reconstruction error.

Figure 3.4 is a basic error system. Further, Figure 3.4 is modified by including the pole-zero model of a signal. The pole-zero model contains the spectral envelope information of a signal. Therefore, signals $S_{HQ2-MSIN}[n']$ and $S_{WB}[n']$ are represented by their respective pole-zero models, as shown in Figure 3.5 [40].

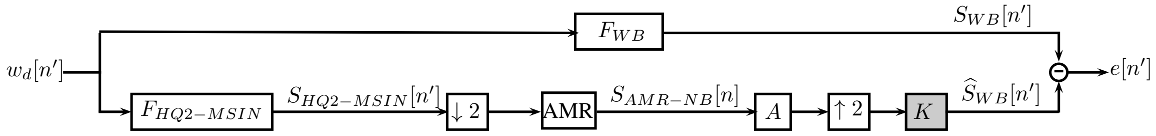

Figure 3.5: Proposed an error system with pole-zero modeling for wideband signal reconstruction.

In Figure 3.5, $S_{HQ2-MSIN}[n']$ and $S_{WB}[n']$ are the outputs of pole-zero models $F_{HQ2-MSIN}$ and F_{WB}, respectively, driven by an input signal $w_d[n']$ with known features (with finite energy, specifically $w_d \in \ell^2(\mathbb{Z}, \mathbb{R}^n)$). In order to obtain a pole-zero model, the number of poles and zeros are fixed as 10, 9 for $F_{HQ2-MSIN}$ and 20, 10 for F_{WB}, respectively. These values are empirically chosen. Signal models $F_{HQ2-MSIN}$ and F_{WB} are then obtained by MATLAB function *prony* based on Prony's method [91]. This function takes the three inputs: signal considered as an impulse response, number of poles, and zeros. The output of the *prony* function is the numerator and denominator coefficients of the signal model. A few poles and zeros of these signal models may lie outside the unit circle. In this case, a minimum phase system is used in the H^∞ optimization problem. This is based on the assumption that the human auditory system is less sensitive to phase information [40]. The poles and zeros lying outside the unit circle are reflected inside the unit circle by inverting their magnitudes without altering the phase for obtaining the minimum phase system [40]. The signal models F_{WB} and $F_{HQ2-MSIN}$ in Figure 3.5 denote the signal models G_1 and G_2 defined in (1.3), respectively. The signal models G_1 and G_2 have the spectral envelope information of the wideband signal (16 kHz) and the narrowband signal (16 kHz), respectively. H_0 is the low pass filter as per the ITU standards, passes the frequency components in the range of 300 Hz to 3400 Hz approximately. H_1 is the p.341 band pass filter, passes the frequency components in the range of 50 Hz to 7000 Hz. H_0 is designed by cascading the MSIN high pass filter and the HQ2 low pass filter.

Problem formulation

The filter K is obtained by minimizing the reconstruction error by the following optimization problem.

Problem 2. Given the signal models $F_{HQ2-MSIN}$, F_{WB}, and filter A, design a stable and causal filter K_{opt} defined as

$$K_{opt} := \arg\min_{K}(\|\mathbb{W}\|_\infty), \tag{3.1}$$

where $\mathbb{W}$ is the discrete error system defined as

$$\mathbb{W} := F_{WB} - K(\uparrow 2)A(\mathrm{AMR})(\downarrow 2)F_{HQ2-MSIN}, \tag{3.2}$$

with input $w_d[n']$ and output $e[n']$ (see Figure 3.5). Here, $\|\mathbb{W}\|_\infty$ denotes the H^∞-norm of the system $\mathbb{W}$, which is defined in (2.1).

Further, a theoretical solution of Problem 2 is obtained using the methods explained in the H^∞ sampled-data control theory [41–45]. To make the problem mathematically tractable, an ideal AMR block (i.e., AMR = 1) has been used only for solving Problem 2. This may result in some modeling errors. However, it is generally advisable to use H^∞-norm in case of modeling errors [52].

Solution of Problem 2

Problem 2 is solved to design an optimal filter K_{opt}. The error system $\mathbb{W}$ is converted into the generalized error system (see Figure B.1) as follows

$$\begin{aligned} G_1(z) =& F_{WB}(z), \\ G_2(z) =& F_{HQ2-MSIN}(z), \\ G_3(z) =& A(z), \\ K_d(z) =& K(z). \end{aligned} \tag{3.3}$$

Further, the solution of Problem 2 is obtained using the solution given in Appendix B. The obtained infinite impulse response (IIR) filter K consists of the narrowband information and high-band information as well. However, only high-band information is required for bandwidth extension. Therefore, the undesired narrowband information present in filter K is suppressed by cascading it with a linear phase FIR high pass filter, which is defined as

$$K_{HPF}(z) = K(z)H_{HPF}(z), \tag{3.4}$$

where K_{HPF} is the synthesis IIR filter used for bandwidth extension. $H_{HPF}(z)$ represents the high pass filter with finite impulse response (FIR). The filter $H_{HPF}(z)$ has a length of 81, which is designed using Matlab command *firls* with setting a cut-off frequency of 3675 Hz (0.45π rad) and subsequently multiplied by the Kaiser window with a shape factor of 2. The filter K_{HPF} is an IIR filter and is represented as a rational transfer function. In order to store the synthesis filter information, the filter K_{HPF} is converted into an FIR filter by truncating higher-order Taylor series coefficients of $K_{HPF}(z)$. The FIR filter length is selected empirically, which is explained in Section 3.2.2. The number of coefficients in the FIR synthesis filter has been fixed to 15, which gives better results overall. In essence, this FIR filter contains the high-band spectral envelope information. This FIR approximation of IIR synthesis filter K_{HPF} is considered as the high-band feature vector $\mathbf{Y_K}$.

3.1.1.3 Gain calculation

The obtained high-band feature vector $\mathbf{Y_K}$ is used in the bandwidth extension technique for estimating the high-band signal $\tilde{S}_{HB}[n']$, as depicted in Figure 3.6. Further, the signal $\tilde{S}_{HB}[n']$ is again passed through a linear phase band pass filter for extracting its desired frequency components between 4 kHz to 7 kHz (approximately). The band pass filter is designed with the specifications: filter order = 40, stopband frequency1 (lower stopband frequency) = 3660 Hz, passband frequency1 (lower passband frequency) = 4340 Hz, passband frequency2 (higher passband frequency) = 7300 Hz, stopband frequency2 (higher stopband frequency) = 7800 Hz, and design method as least square using the Matlab command *designfilt*.

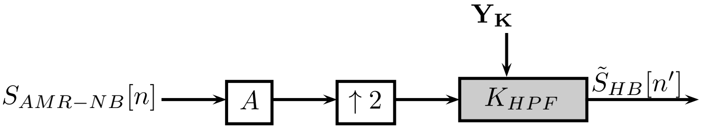

Figure 3.6: Bandwidth extension technique for the AMR coded narrowband signal.

For gain adjustment, the energy of the estimated band pass filtered signal is set equal to the energy of the original band pass filtered signal. As such, the gain factor g is calculated as

$$g = \sqrt{\frac{\sum_{n'=1}^{N} S_{BPF}^2[n']}{\sum_{n'=1}^{N} \tilde{S}_{BPF}^2[n']}}, \tag{3.5}$$

where $S_{BPF}[n']$ is the original band pass filtered signal obtained by band pass filtering of the reference wideband signal $S_{WB}[n']$, $\tilde{S}_{BPF}[n']$ represents the estimated band pass filtered signal derived by band pass filtering of the estimated high-band signal $\tilde{S}_{HB}[n']$, and N is the signal length.

3.1.1.4 Narrowband feature vector extraction

The narrowband envelope information is represented by 16 linear prediction coefficients (LPC), which are calculated for the input signal ($S_{AMR-NB}[n]$) by linear prediction analysis [64]. Additionally, five other features are considered for capturing the input signal characteristics. These features are zero-crossing rate, gradient index, kurtosis, spectral centroid, and normalized relative frame energy [13, 92, 93]. These features are concatenated along with LP coefficients, and the resulting feature vector is represented by $\boldsymbol{x}_i$. Further, temporal characteristics are taken into account by considering adjacent frame's information. The final narrowband feature vector of 63 dimensions is constructed similar to [13]. The narrowband feature vector is composed as

$$\mathbf{X} = \Big[\boldsymbol{x}_i, \quad \boldsymbol{x}_{i+1} - \boldsymbol{x}_{i-1}, \quad \boldsymbol{x}_{i+1} - 2\boldsymbol{x}_i + \boldsymbol{x}_{i-1} \Big],$$

where i, $i-1$, and $i+1$ denote present frame, previous frame, and next frame, respectively.

3.1.1.5 Training of the DNN model

The extracted features $\mathbf{Y_K}, g$, and $\mathbf{X}$ are used to train the DNN model. The narrowband feature vector $\mathbf{X} \in \mathbb{R}^{63}$ is fed to the DNN model as the input. A concatenation of the high-band feature vector $\mathbf{Y_K} \in \mathbb{R}^{15}$ and $\log_{10}$ of squared gain factor g (i.e., $[\mathbf{Y_K}, 2\log_{10} g]$) is taken as the target output for training the DNN model. Here, $2\log_{10} g$ is represented by g_1. Mean squared error is chosen as a loss function for training the DNN model (see DNN-R in [13]). The mean and variance normalization (MVN) has been applied to both the input and output vectors of the DNN model using the statistics obtained for the training set [36].

3.1.2 Artificial bandwidth extension of AMR coded narrowband speech signal

The trained DNN model in Section 3.1.1.5 is used in the artificial bandwidth extension process of the encoded narrowband signal $S_{AMR-NB}[n]$ at the receiver side, as shown in Figure 3.7.

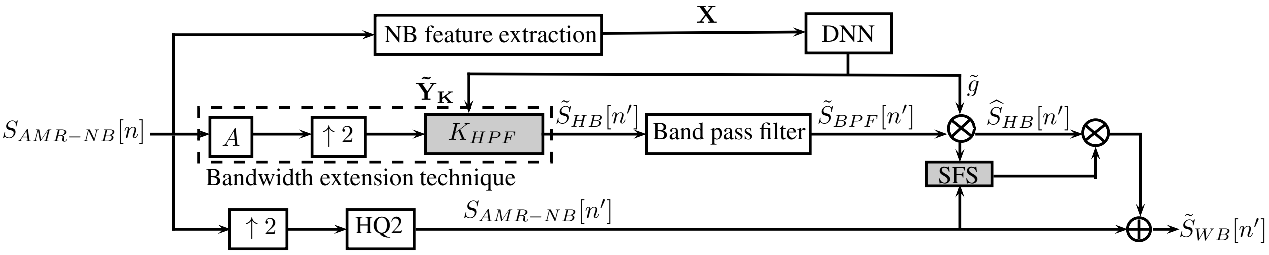

Figure 3.7: Illustration of the artificial bandwidth extension of the AMR coded narrowband signal.

The wideband signal estimation has four main processes. These processes are the narrowband signal reconstruction process, features estimation, high-band signal estimation, and wideband signal estimation, as explained in Sections 3.1.2.1, 3.1.2.2, 3.1.2.3, and 3.1.2.4, respectively.

3.1.2.1 Narrowband signal reconstruction process

The narrowband signal reconstruction process is used to resample the narrowband signal. The AMR coded narrowband signal $S_{AMR-NB}[n]$ sampled at 8 kHz is resampled at 16 kHz. For this, the signal $S_{AMR-NB}[n]$ is upsampled by a factor of 2 and subsequently filtered by the HQ2 low pass filter. This leads to an output signal $S_{AMR-NB}[n']$ sampled at 16 kHz, as shown in Figure 3.7.

3.1.2.2 High-band feature vector and gain factor Estimation

The high-band feature vector and gain factor are estimated using the trained DNN model. For this, the NB feature vector is computed for a given narrowband signal $S_{AMR-NB}[n]$, the same as done in Section 3.1.1.4. MVN is applied to the narrowband feature vector and then fed to the DNN model. A reverse MVN procedure is applied to the DNN output [36]. The output of the DNN model has $[\tilde{\mathbf{Y}}_{\mathbf{K}}, \tilde{g}_1]$, where $\tilde{\mathbf{Y}}_{\mathbf{K}}$ is the estimated high-band feature vector containing the synthesis filter (K_{HPF}) information and $\tilde{g}_1$ contains the corresponding estimated gain factor $(\tilde{g})$ information computed as $\tilde{g} = 10^{(\tilde{g}_1/2)}$.

3.1.2.3 High-band signal estimation

The high-band signal is estimated using the estimated high-band feature vector. The estimated high-band feature vector is used to re-synthesize the high-band signal. For estimating the high-band signal, the analysis filter A is calculated for a given narrowband signal $S_{AMR-NB}[n]$ (see Section 3.1.1.2). The signal $S_{AMR-NB}[n]$ is passed through filter A and then upsampled by a factor of 2. The resulting signal is passed again through the estimated synthesis filter K_{HPF}, which generates a high-band signal $\tilde{S}_{HB}[n']$ (see Figures 3.6, 3.7).

3.1.2.4 Wideband signal estimation

The wideband signal is estimated by adding the resampled narrowband signal and the modified estimated high-band signal obtained using the estimated gain factor and an attenuation factor. The estimated gain factor and attenuation factor are used to set the energy level of the estimated high-band signal. For this, the signal $\tilde{S}_{HB}[n']$ is fed into the band pass filter for extracting the desired frequency components. The obtained signal $\tilde{S}_{BPF}[n']$ is multiplied with the estimated gain factor $\tilde{g}$. This leads to an output signal $\widehat{S}_{HB}[n']$. The spectral floor suppression (SFS) technique [13] is used in the proposed approach. This technique controls the synthesized energy in the high-band frequency range for sounds. For this, the ratio R^{SFS} is computed as

$$R^{\text{SFS}} = \frac{\frac{1}{N/4}\sum_{k=(N/4)+2}^{(N/2)+1} |\widehat{\phi}_{HB}[k]|^2}{\frac{1}{(N/4)+1}\sum_{k=1}^{(N/4)+1} |\widehat{\phi}_{NB}[k]|^2}, \tag{3.6}$$

where $\widehat{\phi}_{HB}[k]$ and $\widehat{\phi}_{NB}[k]$ are the power spectrum density of signals $\widehat{S}_{HB}[n']$ and $S_{AMR-NB}[n']$, respectively. Then an attenuation factor is calculated as

$$d = \min\left\{\frac{d_{\text{high}} - d_{\text{low}}}{\theta^{\text{SFS}}} R^{\text{SFS}} + d_{\text{low}}, d_{\text{high}}\right\} \text{dB}, \tag{3.7}$$

where $d_{\text{high}} = -3$ dB, $d_{\text{low}} = -18$ dB, and $\theta^{\text{SFS}} = 5$. These values have been chosen empirically for the proposed approach.

Finally, the wideband signal $\tilde{S}_{WB}[n']$ is estimated by adding the resampled narrowband signal $S_{AMR-NB}[n']$ and the estimated high-band signal obtained by applying the attenuation factor d on $\widehat{S}_{HB}[n']$ defined as

$$\tilde{S}_{WB}[n'] = S_{AMR-NB}[n'] + 10^{\frac{d}{20}}\widehat{S}_{HB}[n']. \tag{3.8}$$

3.2 Experimental set-up and results

Section 3.2.1 has a description of speech datasets used for evaluating the proposed approach. In Section 3.2.2, experiments are conducted for deciding the number of coefficients in the FIR synthesis filter and DNN topology. Also, spectrogram of a female speech file is analyzed. Objective and subjective measures will be discussed for the proposed approach and compared with two baselines in Sections 3.2.2.2 and 3.2.2.3.

3.2.1 Databases

The proposed approach is evaluated on the TIMIT [73] and RSR15 [74] datasets. The train set of TIMIT dataset is used to train the model, while the test set of TIMIT dataset is considered as a validation set. A new test set is made by taking speech files from the RSR15 dataset. This new test set has the speech files uttered by 4 female and 3 male speakers. Speech files are processed, as explained in Section 3.1.1.1.

3.2.2 Results

Objective and subjective assessments are carried out to analyze the quality of the artificially extended speech signals. For this purpose, objective metrics are chosen the wideband PESQ (perceptual evaluation of speech quality) in terms of the wideband MOS-LQO (mean opinion score listening quality objective) [81, 82], upper-band (4-7 kHz) logarithmic spectral distance (LSD_{UB}), and full-band (0-7 kHz) logarithmic spectral distance (LSD_{FB}) [78]. Artificially extended speech signals are band pass filtered by the standard P.341 filter [2] in the objective assessment. Subjective measure CMOS (comparison mean opinion score) [86] is chosen for examining the speech perceptual quality. The wideband MOS-LQO is used for deciding the high-band feature vector Y_K dimension. The wideband MOS-LQO is measured for the enhanced speech signals belonging to the validation set, which are synthesized by using high-band feature vectors Y_K of different dimensions. We also observe the wideband MOS-LQO by using the SFS technique in the proposed bandwidth extension approach. These analyses are done using the FIR approximation of synthesis filter K_{HPF} directly (oracle filter K_{HPF}). Wideband MOS-LQO values are listed in Table 3.1 without applying the SFS technique ($d = 0$) and in Table 3.2 with applying the SFS technique ($d \neq 0$). It can be observed from Tables 3.1 and 3.2 that the

Table 3.1: Performance evaluation of enhanced speech files belonging to the validation set in the condition of directly using the FIR synthesis filter obtained by truncating the impulse response of IIR synthesis K_{HPF} and without applying the SFS technique ($d = 0$) for ABE

Synthesis Filter Length	0 ($K_{HPF} = 0$)	10	15	20	25	30
MOS-LQO	3.2097	3.3649	3.3996	3.4174	3.3937	3.3846

Table 3.2: Performance evaluation of enhanced speech files belonging to the validation set in the condition of directly using the FIR synthesis filter obtained by truncating the impulse response of IIR synthesis K_{HPF} and applying the SFS technique ($d \neq 0$) for ABE

Synthesis Filter Length	0 ($K_{HPF} = 0$)	10	15	20	25	30
MOS-LQO	3.2097	3.5246	3.5326	3.5310	3.5239	3.5204

SFS technique improves the wideband MOS-LQO value significantly. Also, filter lengths 15 and 20 give almost the same wideband MOS-LQO values and are comparatively better than the other filter lengths for both the cases with and without the SFS technique. Hence, we choose the filter length either 15 or 20 in order to obtain a better wideband MOS-LQO value on the validation set by the DNN model. First, the DNN model is designed for the filter length 15 and then compared with the filter length 20.

3.2.2.1 Architecture of the DNN model

DNN architecture for the proposed HB feature vector along with the gain factor and NB feature vector has been decided experimentally. For this purpose, the batch size (128), the number of maximum epochs (50), momentum (0.9), and the initial learning rate (0.1) have been fixed. The weights and biases are initialized by random values taken from the normal distribution. The normal distribution function is parameterized with zero mean and standard deviation of $u^{-1/2}$, with u being the number of incoming connections of the respective unit. The activation function for the layers has been set to ReLU. For avoiding over-fitting problems, L2-regularization for layer weights has also been employed [70]. In training of the DNN model, the learning rate is fixed according to the validation error. If the validation error is not improved, then the learning rate is changed to half of the previous epoch's learning rate. The minimum learning rate is set to 0.0005. If the learning rate reaches the minimum, then it is not altered. Training of the DNN model is stopped if the validation error does not improve for 5 epochs. Different DNN topologies, obtained by varying the number of hidden layers (N_{HL}) and the number of hidden layer neurons (N_U), have been trained. The mean squared errors computed for predicted outputs of the validation set, generated from different DNN topologies, are computed and compared in Table 3.3. It can be observed from Table 3.3, a topology of 5 N_{HL} and 512 N_U performs best overall. This architecture has been fixed for all the further experiments. Next, this architecture is trained for the filter length of 20. For the validation set, the wideband MOS-LQO value is improved by considering a filter length of 15 rather than 20. Therefore, the number of coefficients in the FIR synthesis filter is taken 15.

Table 3.3: Computation of the mean squared error and standard deviation for the validation set with varying the DNN architecture

Number of hidden-layers	3	3	3	3	4	4	4	4	5	5	**5**	5
Number of Neurons in each hidden-layer	128	256	512	1024	128	256	512	1024	128	256	**512**	1024
Average validation error	0.0508	0.0509	0.0599	0.0642	0.0511	0.0507	0.0506	0.0595	0.0508	0.0572	**0.0504**	0.0683
Standard deviation (10^{-4})	2.4819	1.3565	110.9090	114.4867	1.4697	1.9390	0.4899	114.4867	1.6000	90.3371	**0.4000**	91.2316

Spectrogram of a female speech file taken from the validation set is illustrated in Figure 3.8. We analyzed the spectrogram of different signals obtained at various parts in the proposed ABE framework. Figure 3.8 **(a)**, **(b)**, **(c)**, **(d)**, and **(e)** illustrate spectrograms of the reference wideband speech signal $S_{WB}[n']$, encoded narrowband speech signal $S_{AMR-NB}[n']$ sampled at 16 kHz, extended wideband speech signals using the signals $\tilde{S}_{BPF}[n']$, $\widehat{S}_{HB}[n']$, and $10^{\frac{d}{20}}\widehat{S}_{HB}[n']$ (see Figure 3.7) in the proposed framework using DNN model, respectively. Some fricative sounds (phonemes) such as 's', 'f', and 'sh' are marked in the spectrograms. In Figure 3.8 **(c)**, enhancement is not seen because of not applying the gain factor on the signal $\tilde{S}_{BPF}[n']$. The gain factor is important for perceiving the enhancement. After using the gain factor, enhancement is observed in the spectrogram of signal $\widehat{S}_{HB}[n']$, as shown in Figure 3.8 **(d)**. Some sounds are overestimated in Figure 3.8 **(d)** when compared with Figure 3.8 **(a)**. Hence, the SFS technique is applied on the signal $\widehat{S}_{HB}[n']$, which significantly reduces overestimation, as seen in Figure 3.8 **(e)**. But, energy of the fricatives phonemes is somewhat lessened than energy of the original phonemes. It is happened because of applying the SFS technique, which introduces attenuation in the estimated high-band signal. Further, it is observed that the 's' and 'f' phonemes are reconstructed better than the 'sh' phonemes. A gap or discontinuity at around 4 kHz is observed in Figure 3.8 **(d)** and **(e)** due to using the band-limited narrowband signal. The narrowband signal has frequency contents approximately between 300 Hz to 3400

Hz. This gap in spectral content may degrade perceptual speech quality [10].

3.2.2.2 Objective assessment

The proposed approach is compared with two baselines such as modulation technique [13] and cepstral domain approach [39]. The modulation technique is based on the speech production model. This technique needs the high-band envelope information and high-band residual signal. Therefore, the high-band envelope information is estimated by the linear frequency cepstral coefficients, while the high-band residual signal is obtained by the spectral translation method. The cepstral domain approach estimates the high-band information by finding the high-band magnitude spectrum and high-band phase spectrum. In the cepstral domain approach, the high-band magnitude spectrum is estimated by the linear frequency cepstral coefficients, while the high-band phase spectrum is directly obtained by shifting the phase spectrum of the narrowband signal. The proposed approach is also evaluated using 128 GMMs. Experimental conditions such as window duration, type of window, datasets, and narrowband processing have been fixed in these tests. Wideband MOS-LQO, upper-band logarithmic spectral distance (LSD$_{UB}$), and full-band logarithmic spectral distance (LSD$_{FB}$) are computed for the proposed framework and the baselines on the test set, as arranged in Table 3.4. As it can be observed from Table 3.4, the proposed approach using the DNN model improves all the measures compared to the GMM model. The proposed approach using the DNN model improves by 0.0759 and 0.5482 MOS-LQO values compared to the modulation technique and cepstral domain approach, respectively. The proposed approach using the DNN model improves the LSD$_{UB}$ and LSD$_{FB}$ values when compared to the modulation technique, which may result a better perception of speech sounds. The cepstral domain approach produces the best LSD$_{UB}$ and LSD$_{FB}$ values, which may result

Table 3.4: Performance evaluation on the test set for the proposed approach and the baselines.

Method	Wideband MOS-LQO	LSD$_{UB}$	LSD$_{FB}$
Proposed approach using DNN model	3.3022	17.6657	13.2050
Proposed approach using 128 GMMs model	3.0947	17.9617	13.3834
Modulation technique	3.2263	19.8028	14.4981
Cepstral Domain	2.7540	11.4685	9.7369

a better perception. But, the worst MOS-LQO value is obtained for the cepstral domain approach, which may result the worst speech quality. The proposed approach using the DNN model provides the best MOS-LQO and moderate logarithmic spectral distances (LSD$_{UB}$ and LSD$_{FB}$).

Spectrogram of a female speech file taken from the test set is discussed. Figure 5.9 **(a)**, **(b)**, **(c)**, **(d)**, and **(e)** illustrate spectrogram of the reference speech signal, AMR coded narrowband speech signal sampled at 16 kHz, extended speech signals by the proposed approach, modulation technique, and cepstral domain approach, respectively. It can be observed in Figure 5.9 **(e)**, a pattern like noise is seen in spectrogram of the extended speech signal by the cepstral domain approach. As a result, energy in the estimated high-band region is high, however, this noise affects the speech quality. While this noise is not seen in Figure 5.9 **(c, d)**. Therefore, the speech quality is obtained better for the proposed approach and modulation technique than the cepstral domain approach. Energy in the high-band region of extended speech signal is higher for the proposed approach than the modulation technique. As a result, sounds in extended speech signal are perceived better for the proposed approach than the modulation technique. Some noise may be present in the extended speech signal generated by the proposed approach, however, it does not affect the perception of sounds.

3.2.2.3 Subjective assessment

In a typical telephonic conversation, perceptual quality of the receiving speech signal has been given more priority. For this, subjective assessment is done by following ITU-T P.800 [86, Annex E]. In the subjective assessment, two speech files are compared and scored on the CMOS scale from -3 (much worse) to 3 (much better). Twelve listeners participated in this assessment. They do not have any hearing impairment. Their ages are between 25 to 32. Twelve speech files are taken from the test set for subjective evaluation. CMOS score is calculated for the three conditions in which the artificially extended speech files (enhanced by the proposed approach using the DNN model) are compared to the artificially extended speech files (enhanced by the baselines) and the AMR coded narrowband signals. All these speech files are band pass filtered by the standard P.341 filter [2] and subsequently scaled to an active speech level of -26 dBov [90]. CMOS and 95% confidence interval are listed in Table 3.5 for each test condition. As evident in Table 3.5, the proposed approach improves the AMR coded narrowband speech signal by 1.5208 points. CMOS is improved by 0.5833 and 1.6944 points for the proposed approach in comparison to the modulation technique and cepstral domain approach, respectively. In subjective evaluation, opinions are taken from the listeners. They gave their opinions in terms of noise and word perception. Some listeners prefer the less noisy speech signal, while some prefer the word perception in the speech signal. Noise artifacts are not perceived in enhanced speech files using the modulation technique, however, enhancement in words is not perceived well. It may be due to attenuating the estimated high-band signal. For the cepstral domain approach, noise is perceived higher. For the proposed approach, noise is still perceived, however, it does not affect the perception of words.

Table 3.5: Subjective assessment conducted on the artificially extended speech files belonging to the test set.

Conditions	CMOS	CI_{95}
AMR coded narrowband signal ($S_{AMR-NB}[n']$) vs Proposed approach	1.5208	[1.3064; 1.7352]
Modulation technique vs Proposed approach	0.5833	[0.4613; 0.7053]
Cepstral Domain approach vs Proposed approach	1.6944	[1.5030; 1.8859]

3.3 Conclusion

This work proposes a new ABE approach for speech signals, which uses the H^∞ sampled-data control theory and wideband signal modeling for obtaining a synthesis filter. The H^∞ optimization helps in acquiring the synthesis filter corresponding to a signal model (pole-zero model) and an analysis filter. The synthesis filter contains the high-band spectral envelope information. For adjusting the energy level of the estimated high-band signal, the gain adjustment and the SFS techniques with a custom modification are used in the proposed bandwidth extension approach. A DNN model is used for estimating the synthesis filter in the artificial bandwidth extension framework. The MOS- LQO and CMOS measures are improved by the proposed approach in comparison to the baselines. Later on, a different architecture is proposed in chapter 5 to enhance the results.

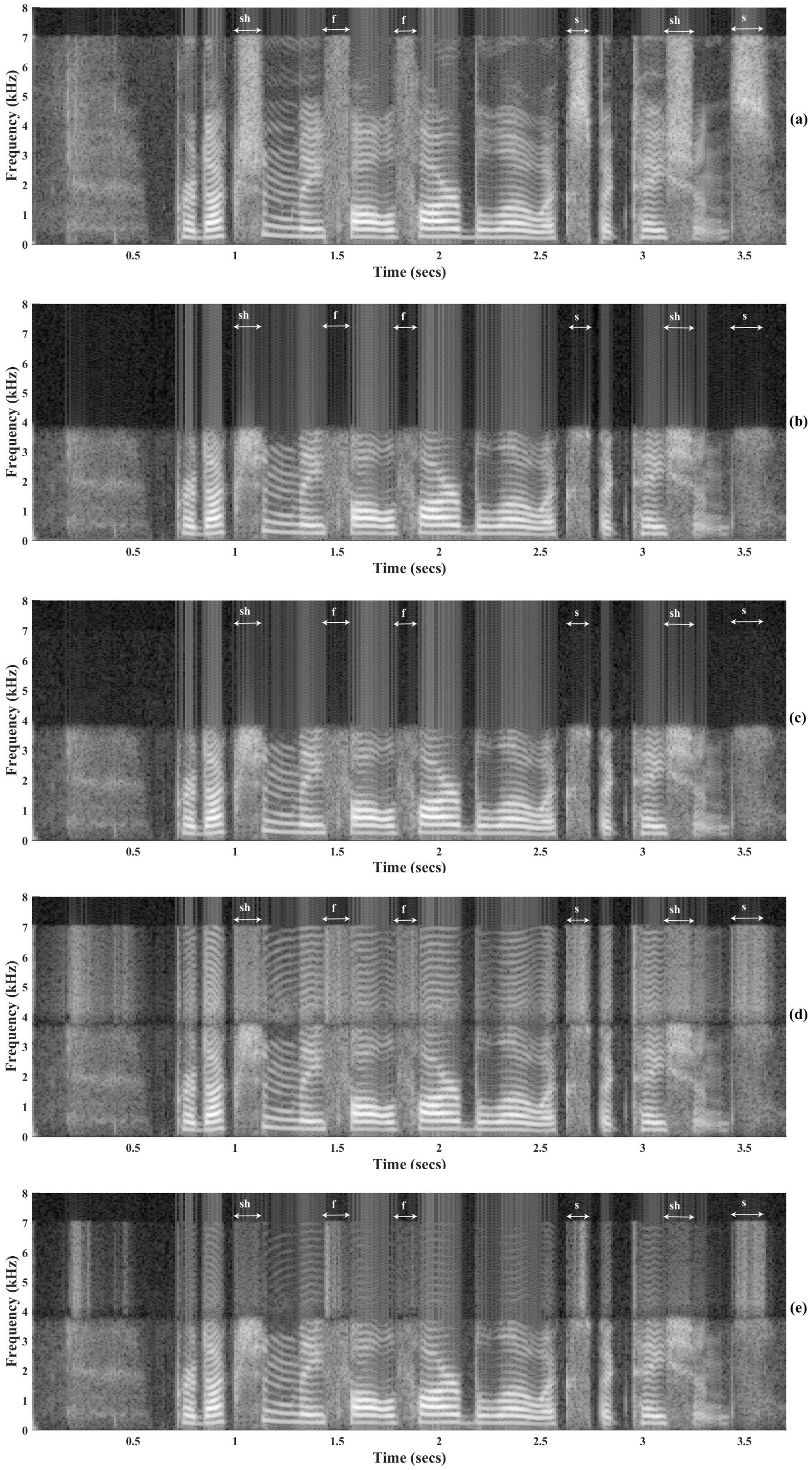

Figure 3.8: Spectrogram of **(a)** reference wideband speech signal of a female speaker, **(b)** AMR coded narrowband signal $S_{AMR-NB}[n']$ sampled at 16 kHz, (**c, d, and e**) extended wideband speech signals using the signals $\tilde{S}_{BPF}[n']$, $\widehat{S}_{HB}[n']$, and $10^{\frac{d}{20}}\widehat{S}_{HB}[n']$ in the proposed framework using DNN model (see Figure 3.7).

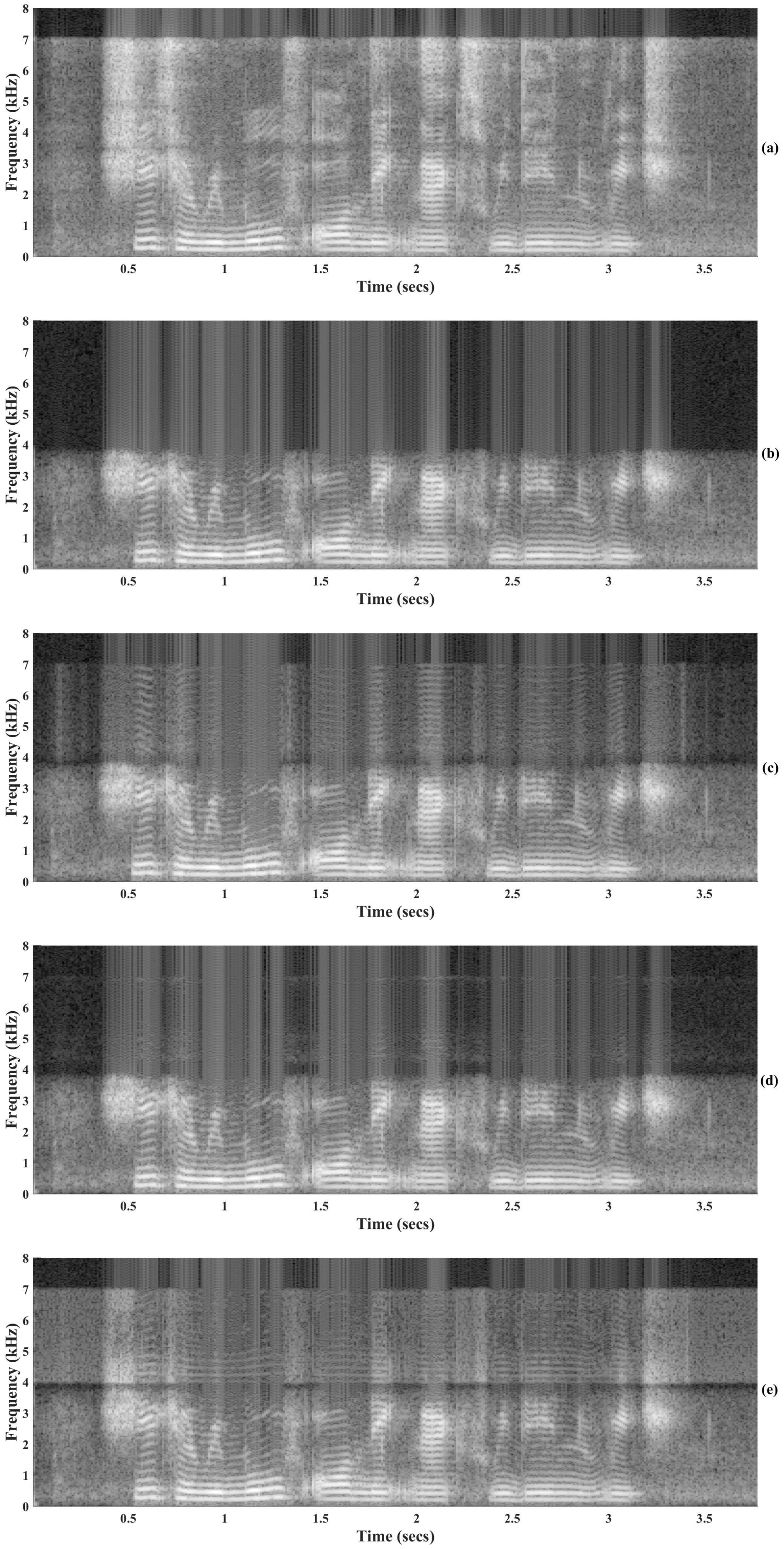

Figure 3.9: Spectrogram of **(a)** reference wideband speech signal of a female speaker, **(b)** AMR coded narrowband signal sampled at 16 kHz, and **(c,d,e)** extended speech signals by the proposed approach, modulation technique, and cepstral domain approach, respectively .

4

Artificial bandwidth extension technique based on the high-band modeling

Contents

4.1 A proposed set-up based on high-band modeling for artificial bandwidth extension of speech signals 61

4.2 Experiment analysis and results 69

4.3 Conclusion 75

In the previous chapter, we have concentrated on the wideband signal model and used a highpass filter to extract the high-band information from the synthesis filter. The major change in this chapter is to consider this highpass filter in the error system for better optimization.

We know that ABE aims to extend the bandwidth of the narrowband (NB) speech signal (up to 4 kHz) to 8 kHz. In this chapter, a new ABE approach is proposed based on the high-band signal modeling. In this context, a signal model is used to represent better the high-band (4-8 kHz) information of a signal. A novel error system is proposed and made by considering the narrowband signal generation process, bandwidth extension process, and reference/true high-band signal generation process. Solution of the error system is obtained using the methods explained in the H^{∞} sampled-data system theory (as in the earlier chapters) [41–45]. The solution of the error system is a synthesis filter for the given signal model. The obtained synthesis filter has high-band (HB) spectral envelope information. The synthesis filter is used in the bandwidth extension process for synthesizing the high-band (4-8 kHz) signal. The DFT concatenation is performed to add the narrowband signal sampled at 16 kHz and the estimated high-band signal sampled at 16 kHz for removing the leaked information from the synthesis filter and non-ideal low pass filter. Gain adjustment is performed on the estimated high-band signal to make its energy equal to the true high-band signal. Non-stationary characteristics of speech signals generate assorted variety in synthesis filters and corresponding gains. For this, a deep neural network (DNN) is trained to estimate the synthesis filter and gain by using only the narrowband information. We analyze the performance of the DNN model on two datasets. Objective and subjective analyses are carried out on these datasets. This chapter is based upon the paper [61].

The rest of the chapter is organized as follows. Section 4.1 has the proposed set-up used for ABE. The proposed set-up uses high-band modeling for extracting high-band information. The proposed approach involves features extraction for designing the DNN model, training of the DNN model, and wideband signal estimation. Section 4.2 contains the objective and subjective analyses. Section 4.3 states a brief conclusion to this work.

4.1 A proposed set-up based on high-band modeling for artificial bandwidth extension of speech signals

This section describes the proposed ABE framework for a narrowband speech signal consisting of frequency components up to 4000 Hz approximately. A basic block diagram for ABE is shown in Figure 1.1. As it can be observed in Figure 1.1, a pre-trained model is needed in advance. The pre-trained model is designed using a database of narrowband features and high-band features. The pre-trained model for the proposed ABE framework is designed in a training block, as shown in Figure 4.1. The training block is elaborated in Section 4.1.1. After designing the pre-trained model, ABE process uses the pre-trained model at the receiver side, as shown in Figure 1.1. As evident from Figure 1.1, the ABE process consists of four main processes: high-band features estimation, NB features extraction, bandwidth extension process, and narrowband signal reconstruction process. These processes play an important role in the estimation of the wideband (WB) signal. Furthermore, the proposed ABE approach uses three additional processes. First process is used for adjusting the energy level of estimated high-band signal. Second is to process the resampled-narrowband signal. Third is to use the DFT concatenation for adding the processed narrowband signal sampled at 16 kHz and estimated high-band signal sampled at 16 kHz. An extension block in Figure 4.1 has a description of the proposed ABE approach for estimating the wideband signal corresponding to the narrowband signal. The extension block is explained in Section 4.1.2.

4.1.1 Training block

The training block consists of three sequential processes: framing process, features extraction process, and modeling process. The framing process is performed to produce stationary speech signals, as explained in Section 4.1.1.1. The features extraction process computes three features: high-band feature vector $\mathbf{Y}_K$, narrowband feature vector $\mathbf{X}$, and gain factor g. The features $\mathbf{Y}_K$, g, and $\mathbf{X}$ are computed in Sections 4.1.1.2, 4.1.1.3, and 4.1.1.4, respectively. The modeling process trains a DNN model using the features. The modeling process is explained in Section 4.1.1.5.

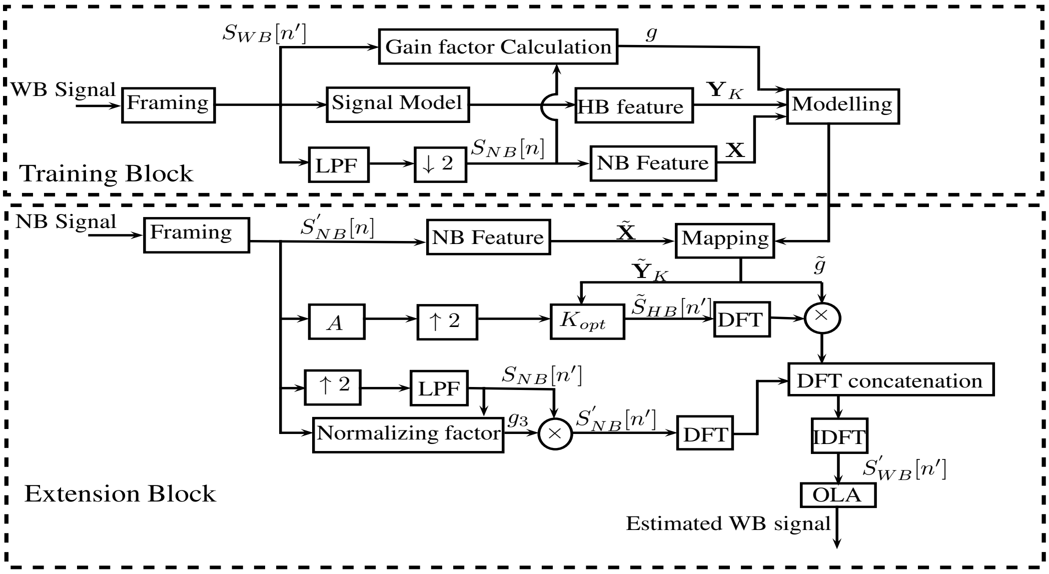

Figure 4.1: Block diagram consists of training of DNN model and artificial bandwidth extension of the narrowband signal.

4.1.1.1 Framing

Speech signals are segmented into frames, and these frames are considered as stationary signals. Here, speech signals are windowed into frames of 25 ms duration with 50% overlapping between adjoining frames using the Hamming window.

4.1.1.2 High-band feature vector extraction

The high-band feature vector $\mathbf{Y_K}$ contains information of the proposed synthesis filter, which is used in the bandwidth extension process. The bandwidth extension process is employed on a stationary NB speech signal (NB frame) $S_{NB}[n]$ for estimating the corresponding HB signal $\tilde{S}_{HB}[n']$, as shown in Figure 4.2, where n and n' denote 8 kHz and 16 kHz sample index, respectively. In Figure 4.2, A is a linear discrete time-invariant (LDTI) LP analysis filter, ↑ 2 is an ideal upsampler with upsampling factor 2, and filter K is an LDTI synthesis filter. The

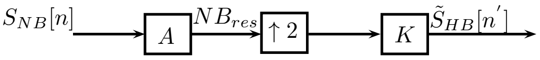

Figure 4.2: Bandwidth extension process applied to a stationary narrowband signal in order to estimate the corresponding high-band signal.

transfer function of filter A is the inverse transfer function of an all-pole filter. The all-pole filter of order 11 is found using the signal $S_{NB}[n]$ with the help of linear prediction analysis (see Figure 4.2) [5]. An NB residue signal (NB_{res}) is the response of filter A driven by the signal $S_{NB}[n]$ [4]. The WB residue signal is an upsampled NB residue signal by a factor of 2. It is fed into the synthesis filter K in order to estimate the high-band signal $\tilde{S}_{HB}[n']$.

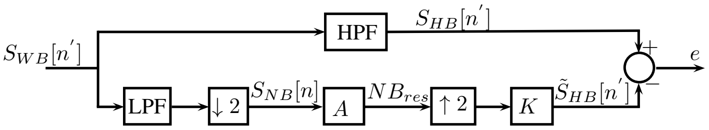

Figure 4.3: An error system set-up.

Here, our primary focus is to design the synthesis filter K in order to estimate the HB information $\tilde{S}_{HB}[n']$ related to the NB signal $S_{NB}[n]$. It can be done by considering the NB signal $S_{NB}[n]$ generation process, bandwidth extension process, and HB signal generation process. For this, an error system is made, as depicted in Figure 4.3. In Figure 4.3, HPF is a non-causal FIR high pass filter (HPF), which produces the true/original high-band signal $S_{HB}[n']$. In high-band signal generation process, the signal $S_{HB}[n']$ is generated by high pass filtering of the the original wideband signal $S_{WB}[n']$. In this chapter, we focus on reconstruction of $S_{HB}[n']$, which contain information about the high-band frequencies. This is justified as narrowband information is available at the receiver end and we can utilize it as it is. LPF is a non-causal FIR low pass filter. In the narrowband signal generation process, the narrowband signal $S_{NB}[n]$ is generated by passing the wideband signal $S_{WB}[n']$ through the LPF and subsequent downsampling by a factor of 2 at the transmitter side. The synthesis filter K is designed for minimizing the reconstruction error. We use a system norm to measure the reconstruction error [59]. In Figure 4.3, $e = S_{HB}[n'] - \tilde{S}_{HB}[n']$, i.e., the error between the true HB signal and estimated HB signal. To minimize the error, it is beneficial to extract prior information associated with the wideband speech signal $S_{WB}[n']$. A signal model F is used to represent the prior information of the signal $S_{WB}[n']$. It is taken into account for Figure 4.3, and the resulting set-up is shown in Figure 4.4. Here, H_0 and H_1 denote the LPF and HPF, respectively. The signal $S_{WB}[n']$ is the output of the signal model F driven by an input w with known features (with finite energy,

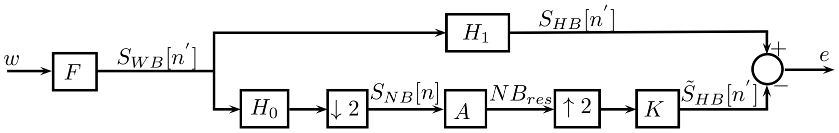

Figure 4.4: A proposed architecture of the error system set-up for estimating the high-band signal.

specifically $w \in \ell^2(\mathbb{Z}, \mathbb{R}^n)$). $F(z)$ representing the rational transfer function of F, is assumed to be a stable and causal transfer function. This model can be obtained by Prony's method, available in MATLAB [62, 63]. The obtained model is causal but may be unstable. Hence, it is converted into a stable model by inverting its unstable poles. This does not affect the magnitude spectrum of $F(z)$ but changes the phase [40]. The human auditory system is less sensitive to phase of the speech signal [40]. Further, the number of poles and zeros in the signal model was empirically calculated for each frame in such a way that the minimizes the error. In Figure 4.4, H_1F and H_0F denote the signal models G_1 and G_2 defined in (1.3), respectively. Signal models G_1 and G_2 have the spectral envelope information of the high-band signal (16 kHz) and narrowband signal (16 kHz), respectively. When compared to (1.3), we can easily see that $G_1 = H_1F$ and $G_2 = H_0F$ in Figure 4.4. The signal of interest is the high-band signal $S_{WB}[n']$. Hence, high-band modeling is performed.

Problem formulation

We solve the following optimization problem for designing an optimal $K(z)$.

Problem 3. *Given a stable and causal transfer function $F(z)$, two non-causal FIR filters $H_0(z)$ and $H_1(z)$, design an optimal stable and causal synthesis filter K_{opt} defined as*

$$K_{opt} := \arg\min_{K}(\|\mathbb{P}\|_\infty), \tag{4.1}$$

where $\mathbb{P} := H_1F - K(\uparrow 2)A(\downarrow 2)H_0F$. $\mathbb{P}$ maps w to e in Figure 4.4. Here, $\|\mathbb{P}\|_\infty$ represents the H^∞-norm of the system $\mathbb{P}$.

Solution of Problem 3

Problem 3 is solved to obtain an optimal filter K_{opt}. Filters H_0 and H_1 present in system $\mathbb{P}$ are non-causal systems. Thereby, system $\mathbb{P}$ is a non-causal system. Hence, this system needs to be converted into a causal system for obtaining the solution using the solution given in

Appendix B. FIR filters H_0 and H_1 have a relation to each other, which can be written in z-domain [48]

$$H_1(z) = H_0(-z). \tag{4.2}$$

Consider the FIR filter $H_0(z)$

$$\begin{aligned} H_0(z) =& a_Q z^{-Q} + .. + a_1 z^{-1} + a_0 + a_1 z^1 + ... + a_Q z^Q, \\ =& z^Q H_a(z), \end{aligned} \tag{4.3}$$

with $H_a(z) := (a_Q z^{-2Q} + .. + a_1 z^{-(Q+1)} + a_0 z^{-Q} + a_1 z^{-(Q-1)} + ... + a_Q)$ with $a_i \in \mathbb{R}$ and Q can be assumed as an even integer number without the loss of generality. The filter $H_1(z)$ can be obtained by substituting (4.3) into (4.2), i.e.,

$$H_1(z) = z^Q H_a(-z) \tag{4.4}$$

Next, we replace $H_0(z)$ by (4.3) and $H_1(z)$ by (4.4) in the system $\mathbb{P}$ as

$$\begin{aligned} \mathbb{P}(z) =& z^Q H_a(-z)F(z) - K(z)(\uparrow 2)A(z)(\downarrow 2)z^Q H_a(z)F(z), \\ =& z^Q(F_b(z) - K(z)(\uparrow 2)A(z)(\downarrow 2)F_a(z)), \end{aligned} \tag{4.5}$$

where $F_b(z) := H_a(-z)F(z)$ and $F_a(z) := H_a(z)F(z)$. Further, the system $\mathbb{P}$ is transformed into a causal system by delaying its response to Q samples as

$$\mathcal{P} = z^{-Q}\mathbb{P}, \tag{4.6}$$

where the system $\mathcal{P}$ is a causal system. The H^∞-norm of the system $\mathcal{P}$ is equivalent to the original system $\mathbb{P}$ due to the fact that the delaying process does not change the H^∞-norm of the system [1]. Further, the solution of (4.6) is obtained using the solution given in Appendix B wherein the system $\mathcal{P}$ is converted into the generalized error system (see Figure B.1) as follows

$$\begin{aligned} G_1(z) &= F_b(z), \\ G_2(z) &= F_a(z), \\ G_3(z) &= A(z), \\ K_d(z) &= K(z). \end{aligned} \tag{4.7}$$

The obtained optimal filter K (K_{opt}) contains the high-band information of the wideband signal. Filter K has infinite impulse response (IIR). So, it is converted into an approximate FIR filter by truncating the Taylor series of K at the origin, which is taken as the high-band feature vector $\mathbf{Y_K}$. The number of coefficients in the FIR filter is taken 20 experimentally, as explained in Section 4.2.

4.1.1.3 Gain factor calculation

The estimated HB signal $\tilde{S}_{HB}[n']$ can have different energy level from the corresponding true high-band information present in $S_{WB}[n']$. Hence, a gain g factor is calculated as

$$g = \sqrt{\frac{\sum_{k=\frac{N}{4}+1}^{\frac{N}{2}} |S_{WB}[k]|^2}{\sum_{k=\frac{N}{4}+1}^{\frac{N}{2}} |\tilde{S}_{HB}[k]|^2}}, \tag{4.8}$$

where, $S_{WB}[k]$ and $\tilde{S}_{HB}[k]$ are the discrete Fourier transforms (DFTs) of $S_{WB}[n']$ and $\tilde{S}_{HB}[n']$, respectively. k represents the frequency bin index. N is the number of samples in a signal sampled at 16 kHz.

4.1.1.4 Narrowband feature vector extraction

The narrowband information is represented by line spectral frequencies (LSF) [65], which are computed using the narrowband signal $S_{NB}[n]$. The dimension of the NB feature vector $\mathbf{X}$ is fixed to 11.

4.1.1.5 Modeling

In modeling process, a DNN model is trained, which is taken as the pre-trained model. The DNN model is structured using the NB features, high-band features, and gain factor. The NB feature vector $\mathbf{X} \in \mathbb{R}^{11}$ is taken as the input of the DNN model. A vector $\mathbf{W} \in \mathbb{R}^{21}$ is obtained by concatenating the high-band feature vector $\mathbf{Y}_K \in \mathbb{R}^{20}$ and gain factor $g \in \mathbb{R}$, i.e., $\mathbf{W} = [\mathbf{Y}_K, g]$. $\mathbf{W}$ is taken as the output of the DNN model. The mean and variance normalization (MVN) has been applied to both the input and output vectors of the DNN model using the statistics obtained for the training set [36].

4.1.2 Extension block

In the extension block, the pre-trained DNN model designed in Section 4.1.1.5 is used for the artificial bandwidth extension of the narrowband signal. The wideband signal is synthesized using the four processes: narrowband signal process, mapping process for estimating the high-band feature vector and gain factor, high-band signal estimation, and wideband signal estimation using the DFT concatenation, as explained in Sections 4.1.2.1, 4.1.2.2, 4.1.2.3, and 4.1.2.4, respectively.

4.1.2.1 Narrowband signal process

The NB signal $S'_{NB}[n]$ sampled at 8 kHz is converted into 16 kHz sampling rate signal. For this, $S'_{NB}[n]$ is passed through the upsampler $\uparrow 2$ followed by the LPF. The resulting NB signal $S_{NB}[n']$ is multiplied by a normalizing factor g_3 calculated as

$$g_3 = \frac{\max(S'_{NB}[n])}{\max(S_{NB}[n'])},$$

and the resulting signal is the final NB signal $S'_{NB}[n']$ (see Figure 4.1). This factor makes the peak value of $S'_{NB}[n']$ equal to the peak value of $S'_{NB}[n]$, which improves the results as seen later in Section 4.2.

4.1.2.2 Mapping process

We have the only narrowband signal. So, we compute the NB feature vector $\tilde{\mathbf{X}}$ using the given stationary NB signal $S'_{NB}[n]$, as done in Section 4.1.1.4. In the mapping process, the NB feature vector $\tilde{\mathbf{X}}$ is fed into the pre-trained DNN model, and the resulting output of DNN gives the estimated feature vector $\tilde{\mathbf{W}} = [\tilde{\mathbf{Y}}_K, \tilde{g}]$.

4.1.2.3 Estimation of the high-band signal

The estimated HB feature vector $\tilde{\mathbf{Y}}_K$ has the filter coefficients of filter K_{opt}, which is used in the estimation of HB signal $\tilde{S}_{HB}[n']$ (see Section 4.1.1.2 and Figure 4.1).

4.1.2.4 Wideband signal estimation using the DFT concatenation

We are not going to add $S'_{NB}[n']$ and $\tilde{S}_{HB}[n']$ signals directly for estimating the WB signal. Because both signals are not fully ideal, i.e., $S'_{NB}[n']$ and $\tilde{S}_{HB}[n']$ have some HB information and NB information, respectively. Therefore, we are going to add them in the frequency domain. For this, N-point DFTs of $S'_{NB}[n']$ and $\tilde{S}_{HB}[n']$ are computed and denoted by $S'_{NB}[k]$ and $\tilde{S}_{HB}[k]$, respectively. Then, the initial $\frac{N}{2}+1$ DFT points of $S'_{WB}[k]$ are obtained by considering the initial $\frac{N}{4}+1$ DFT points of $S'_{NB}[k]$ and the $\frac{N}{4}+1$ to $\frac{N}{2}$ DFT points of $\tilde{S}_{HB}[k]$, i.e.,

$$S'_{WB}[k] = \left\{ \begin{array}{ll} S'_{NB}[k], & k = 0, ..., \frac{N}{4} \\ \tilde{g}\tilde{S}_{HB}[k]. & k = \frac{N}{4}+1, ..\frac{N}{2} \end{array} \right\}$$

Rest DFT points between $\frac{N}{2}+1$ to $N-1$ of $S'_{WB}[k]$ are obtained by its initial $\frac{N}{2}+1$ DFT points as

$$S'_{WB}[k] = S'_{WB}\left[\frac{N}{2} - i\right], \begin{array}{c} i = 1, 2, ..\frac{N}{2} - 1, \\ k = \frac{N}{2} + i. \end{array}$$

Here, we call this entire process as DFT concatenation. The inverse DFT (IDFT) $S'_{WB}[k]$ is producing the estimated wideband signal $S'_{WB}[n]$. The DFT concatenation discards the leaked information from the non-ideal low pass filter and filter K_{opt}. Afterward, the full wideband speech signal is obtained by using the overlap add method (OLA) [71] from the estimated wideband signals of different frames.

4.2 Experiment analysis and results

In this section, we performed experiments to establish correctness and effectiveness of the proposed approach. Section 4.2.1 has a description of speech datasets and parameters used for evaluating the proposed approach. In Section 4.2.2, objective metrics are discussed, experiments are conducted for deciding the number of coefficients in the FIR synthesis filter, and performances are analyzed at various parts in the proposed framework of ABE. In Section 4.2.3, experiments are conducted for deciding the DNN topology. The comparison of the performances between the proposed approach and the existing approaches has been discussed in Section 4.2.4. In Section 4.2.5, a subjective assessment is carried out to check the speech perceptual quality.

4.2.1 Databases and parameters

The proposed approach is evaluated on the two datasets: TIMIT dataset [73] and RSR15 dataset [74]. The train set of TIMIT dataset is used to extract training features for training the DNN model, while the test set of TIMIT dataset is considered as a validation set for deciding the DNN architecture. A new test set is constructed using the RSR15 dataset. This new test set has the speech files uttered by 4 female and 3 male speakers. The DNN model is tested on the test set.

In Figure 4.1, filters LPF and HPF are needed. These filters are the non-causal FIR filters, as considered earlier. Firstly, a causal FIR LPF of length 41 is constructed by following the two sequential processes: one is to generate an FIR LPF filter using the command *firls* in MATLAB, and the second is to multiply the obtained filter with the *Kaiser* window in MATLAB [63]. The non-causal FIR LPF filter H_0 (symmetric about the y-axis) is then obtained by advancing the impulse response of obtained causal FIR LPF filter to 20 samples. The filter H_1 is designed directly from the filter H_0 by following (4.2).

4.2.2 Objective analysis

In this work, upper-band (4-8 kHz) logarithmic spectral distance (LSD$_{UB}$), full-band (0-8 kHz) logarithmic spectral distance (LSD$_{FB}$) [39], narrowband MOS-LQO (mean opinion score listening quality objective) [79, 80], and wideband MOS-LQO [81, 82] as objective measures are taken for examining the quality of artificially extended speech signals. The mathematical formulations of these measures are presented in Appendix C.

Table 4.1: Performance evaluation on the validation set in condition of direct use of FIR filter obtained by truncating the impulse response of IIR K_{opt} in Figure 4.1 for ABE.

Filter length	11	15	20	25	30
LSD$_{FB}$	6.2814	6.2951	6.3059	6.3199	6.3286
LSD$_{UB}$	8.3901	8.4132	8.4312	8.4518	8.4651
Narrowband MOS-LQO	4.5200	4.5200	4.5201	4.5201	4.5201
Wideband MOS-LQO	3.5533	3.5480	3.5609	3.5469	3.5563

Further, we convert the IIR filter K_{opt} into an approximate FIR filter by using the Taylor series truncation method. For deciding the number of coefficients in FIR, we evaluate the objective measures produced by the FIR filters of different lengths on the validation set, as arranged in Table 4.1. Here, we choose the filter length 20 because of obtaining the moderate measures.

The objective measures are analyzed by including the normalizing factor g_3, proposed filter with the DFT concatenation, and gain factor (see Section 4.1) in the proposed framework. For this, we use the proposed FIR filter K_{opt} directly (oracle K_{opt}) in Figure 4.1 for estimating the WB signal. Then, WB signals sampled at 16 kHz are estimated with the help of three different outputs such as $S_{NB}[n']$, $S'_{NB}[n'] = g_3 S_{NB}[n']$, and $S'_{WB}[n']$ in Figure 4.1. This is done by applying the OLA method directly on them for computing the corresponding WB signal. The objective measures are computed in Table 4.2 on the validation set for these three conditions. As it can be observed in Table 4.2, the measures LSD$_{FB}$, LSD$_{UB}$, and wideband MOS-LQO are improved for the signal $S'_{NB}[n']$, while the narrowband MOS-LQO is slightly degraded in comparison to the signal $S_{NB}[n']$. After synthesizing the HB signal $\tilde{S}_{HB}[n']$ using

Table 4.2: Performance evaluation on the validation set for the signals $S_{NB}[n']$, $S'_{NB}[n'] = g_3 S_{NB}[n']$, $S'_{WB}[n']$ in Figure 4.1 for ABE.

Conditions	$S_{NB}[n']$	$S'_{NB}[n']$	$S'_{WB}[n']$
LSD$_{FB}$	13.7871	10.0571	6.3059
LSD$_{UB}$	17.7817	13.8457	8.4312
Narrowband MOS-LQO	4.5417	4.5201	4.5201
Wideband MOS-LQO	3.8670	3.8822	3.5609

the oracle FIR filter K_{opt}, the wideband signal is estimated using the DFT concatenation along with gain, which improves the LSD$_{FB}$ by 3.7512 dB, LSD$_{UB}$ by 5.4145, and wideband MOS-LQO by 0.3213 points and maintains the same narrowband MOS-LQO when compared to the signal $S'_{NB}[n']$. The synthesis filter consists of the spectral envelope information of a signal, and the gain factor adjusts the energy of the synthesized high-band signal. Therefore, the LSD is improved using the synthesis filter and the gain factor. The wideband MOS-LQO value is degraded because of the presence of noise artifacts in the synthesized wideband signal. Further, we evaluate the performances of the DNN model.

4.2.3 DNN model performance

Firstly, experiments are performed to finalize the DNN architecture. Hyper-parameters such as learning rate and mini-batch size are decided empirically. These parameters are optimized as per the best performance on the validation set. For this, the number of hidden layers (N_{HL}) and the number of units (N_U) in hidden layers are selected 3 and 512, respectively. Also, we fixed Adamax (adaptive moment estimation based on the infinity norm) [83] optimizer, Relu activation function in hidden layers, and linear activation function in the output layer. For Adamax optimizer, decay rates $\beta 1$ for the first-moment estimate and $\beta 2$ for the second-moment estimate are fixed to 0.9 and 0.999, respectively. Batch normalization, early stopping criteria, and L2-regularization are used in designing the DNN model. In addition, the mean-variance normalization [36] is applied to the feature vectors of the training set, validation set, and test set by using the statistics obtained for the training set. Next, the learning rate is varied in the range of 0.5 to 0.001 and the mini-batch size is varied in the range of 128 to 1024. Maximum epochs are set to 50. DNN models are designed for different learning rates and mini-batch sizes, and their performances are analyzed on the validation set, as shown in Table 4.3.

Table 4.3: Objective analysis on the validation set by varying the learning rate and mini-batch size for the fixed DNN topology with 3 N_{HL} and 512 N_U and Relu activation function in hidden layers.

Learning rate	Batch size	LSD$_{FB}$	LSD$_{UB}$	Narrowband MOS-LQO	Wideband MOS-LQO
0.5	512	6.8344	9.2136	4.5201	3.0494
0.1	512	6.8022	9.1659	4.5201	3.0918
0.01	512	6.8602	9.2516	4.5201	2.9815
0.001	512	7.4642	9.6443	4.5201	2.6681
0.1	128	6.7867	9.1426	4.5201	3.1054
0.1	256	6.8074	9.1736	4.5201	3.0834
0.1	**768**	6.7819	9.1359	4.5201	3.1154
0.1	1024	6.8002	9.1626	4.5201	3.0947

It can be observed that DNN model trained using 0.1 (learning rate) and 768 (mini-batch size) performs better. Therefore, these values are fixed in further experiments. Further, different DNN models are designed by changing the number of hidden layers (N_{HL}) and the number of units (N_U) in hidden layers. Then, the objective analysis is done on the validation set in Table 4.4. In Table 4.4, the narrowband MOS-LQO is not affected by any architecture, i.e., narrowband is not affected in extension by using different estimated synthesis filters due to the DFT concatenation. An architecture with 7 hidden layers and 256 neurons in each layer

Table 4.4: Objective analysis on the validation set by varying the number of hidden layers (N_{HL}) and the number of units (N_U) in hidden layer for the fixed batch size 768, and Relu activation function in hidden layers.

N_{HL}	N_U	LSD$_{FB}$	LSD$_{UB}$	Narrowband MOS-LQO	Wideband MOS-LQO
128	5	6.7894	9.1506	4.5201	3.1122
128	6	6.7617	9.1036	4.5201	3.1309
128	7	6.7762	9.1276	4.5201	3.1159
256	5	6.8012	9.1640	4.5201	3.0753
256	6	6.7578	9.1040	4.5201	3.1328
256	**7**	6.7461	9.0838	4.5201	3.1629
256	8	6.7655	9.1126	4.5201	3.1472
512	4	6.7764	9.1276	4.5201	3.1253
512	5	6.7731	9.1556	4.5201	3.1016

yields the better LSD$_{FB}$, LSD$_{UB}$, and wideband MOS-LQO. This architecture is chosen as the optimal DNN architecture.

4.2.4 Performances comparison

The proposed approach is compared with the existing approaches keeping the same experimental conditions as LPF, HPF, dimension of HB feature vector, DNN architecture (7 hidden layers and 256 neurons in each hidden layer), dataset and NB signal processing. Two recently reported current works such as modulation technique [13] with slight modification and cepstral domain approach [39] are included for comparison. Gain for the modulation technique is calculated by following [55], and the cepstrum features are used for representing the NB information as well as the HB spectral envelope information. The NB feature vector and HB feature vector in the cepstral domain approach contain the NB magnitude spectrum representing the NB information and the cepstral coefficients representing the HB magnitude spectrum [39], respectively. Objective measures are arranged in Table 4.5 for the proposed approach and the existing methods using the same DNN model. The LSD measure is improved by the proposed approach rather than the existing approaches, as viewed in Table 4.5. The proposed synthesis filter has more magnitude information. Therefore, LSD$_{FB}$ and LSD$_{UB}$ measures are improved by the proposed approach. Word perception is higher for the proposed approach due to better LSD$_{FB}$ and LSD$_{UB}$. The narrowband MOS-LQO is obtained approximately the same for the proposed approach and cepstral domain approach and improved slightly for the proposed approach in comparison to the modulation technique. The narrowband region is somewhat affected by the estimated high-band signal in the modulation technique. Therefore, the narrowband MOS-LQO is slightly degraded. The wideband MOS-LQO value is obtained better by

Table 4.5: Objective analysis on the test set for the proposed approach and the existing approaches.

Method	LSD$_{FB}$	LSD$_{UB}$	Narrowband MOS-LQO	Wideband MOS-LQO
Proposed approach	7.9792	10.7610	4.2602	2.8439
Modulation technique	8.3985	11.2912	4.2292	2.9021
Cepstral Domain approach	9.8444	13.4141	4.2601	3.1718

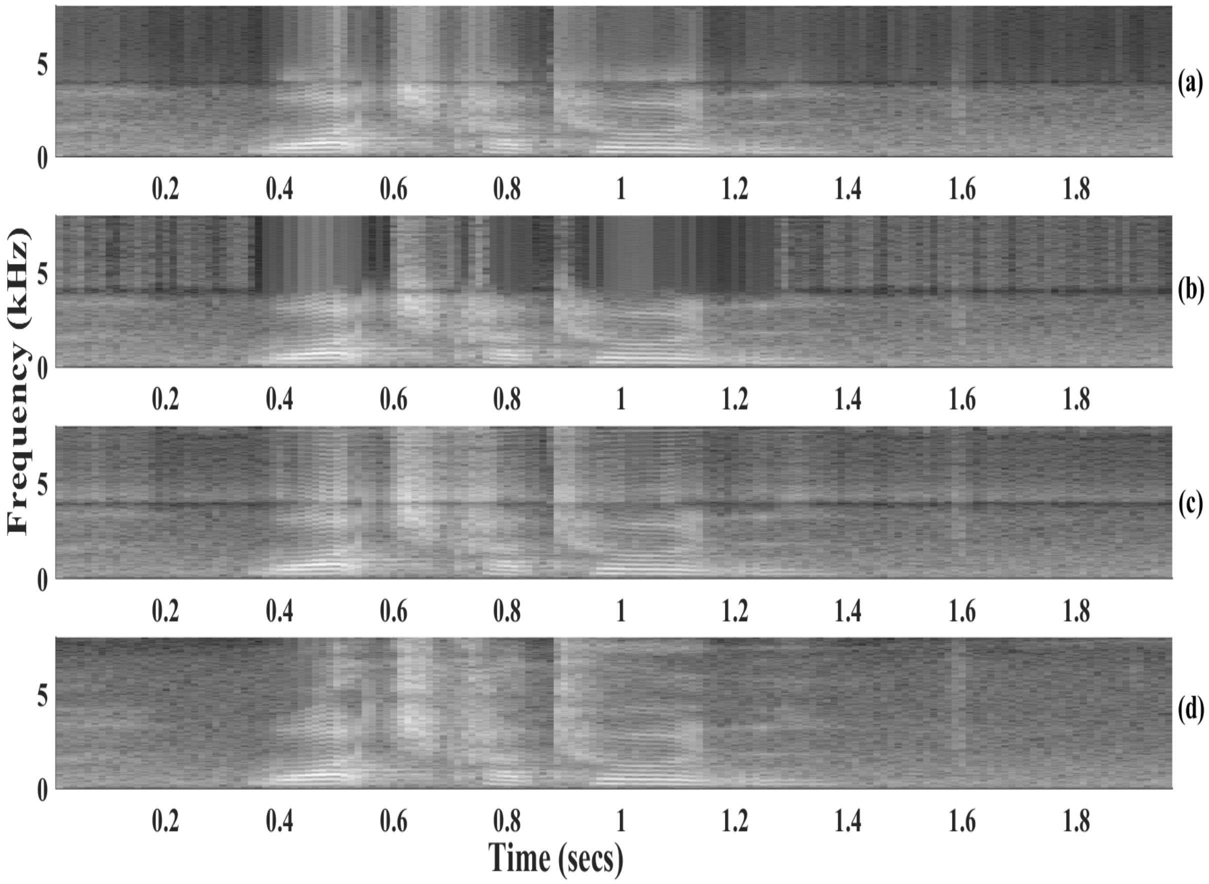

Figure 4.5: Spectrogram of: **(a)** artificially extended speech signal by the cepstral domain approach using DNN model, **(b)** artificially extended speech signal by the **modulation technique** using DNN model, **(c)** artificially extended speech signal by the **proposed approach** using DNN model, and **(d)** original WB signal

the cepstral domain approach than the modulation technique and proposed approach. In the cepstral domain approach, noise artifacts in enhanced speech signals are introduced less than the proposed approach and modulation technique.

Moreover, we visualize the spectrogram of the artificially extended speech signal by using the same DNN model for the proposed approach and the existing approaches. In Figure 4.5, the spectrogram of an artificially extended speech signal is more close to its original spectrogram for the proposed approach than the existing approaches by using the same DNN model.

4.2.5 Subjective listening test

The subjective listening test is also done to check the speech quality based on the perception by following the ITU-T P.800 [86, Annex E]. In this test, speech signals obtained by two conditions are compared with each other by listening, and the rating is done on the comparison MOS (CMOS) scale from -3 (much worse) to 3 (much better). Twelve speakers do it for 20 speech signals taken from the test set. As evident in Table 4.5, the modulation technique performs well in comparison to the cepstral domain approach. Therefore, the CMOS is measured for the proposed approach with respect to the modulation technique where the CMOS has obtained 0.80 points for the proposed approach.

4.3 Conclusion

In this chapter, a new ABE architecture is proposed, which is based on the high-band modeling and H^∞ optimization. An optimal synthesis filter K_{opt} is designed with the help of the H^∞ optimization by using a signal model and an analysis filter. The synthesis filter is used in the estimation of the high-band signal. The filter K_{opt} is an IIR filter. It is converted into an FIR filter, and the resulting filter coefficients are represented by the HB feature vector. Besides, the DFT concatenation is preferred over the time addition for combining the estimated high-band signal and narrowband signal. It removes the leaked NB information in the estimated high-band signal and the high-band information in the narrowband signal. DNN model is used to estimate the synthesis filter information and gain for a given NB feature vector. We obtained the best LSD$_{FB}$ and LSD$_{UB}$ measures by the proposed approach in comparison to the existing approaches. The subjective measure is improved by the proposed approach when compared to the modulation technique.

5

Artificial bandwidth extension technique based on the mapped high-band modeling

Contents

5.1 Proposed framework for the artificial bandwidth extension of speech signal 78

5.2 Speech databases, measures, and results analysis 89

5.3 Objective comparison with the previous schemes 95

5.4 Conclusion 97

In this chapter, we extend the previous ABE approach proposed in chapter 4. Also, the proposed approach is compatible with a practical transmitter set-up. The proposed ABE approach is based on the mapped high-band signal modeling (shifting the high-band frequency in the narrowband) and H^{∞} optimization. This is because the H^{∞} optimization works well in the narrowband (similar idea was used in a different context [94]). Further, an error system is proposed for minimizing error in the case of mapped high-band signal modeling. The error system is built up by combining the narrowband signal generating process, bandwidth extension process, and reference signal generating process. The reference signal is then the mapped high-band signal or band pass shifted signal, which is the original high-frequency components shifted into the narrowband region. A gain factor corresponding to the synthesis filter is computed and used for adjusting the energy levels of the estimated high-frequency components. Speech signals have time-varying characteristics. Therefore, several synthesis filters and corresponding gains are needed for constructing the whole speech signal. Hence, two different deep neural networks (DNNs) are designed for estimating the synthesis filter information and gain factor. We design separate DNN models for modeling the synthesis filter and the gain factor. In addition, the gain factor is computed and modeled in such a way that the gain factor reduces the performance loss obtained due to error in the predicted synthesis filter.

The rest of the chapter is organized as follows: Section 5.1 contains details of the proposed ABE framework. The proposed framework considers speech file operations, features derivation process, training of deep neural networks, and extension of the encoded narrowband signal. Section 5.2 has details of the databases used for analyzing the proposed approach, measures used for evaluating the proposed approach, and results analysis. In Section 5.3, the proposed scheme is compared to the previous schemes. Section 5.4 concludes the proposed approach.

5.1 Proposed framework for the artificial bandwidth extension of speech signal

A basic block diagram for ABE is shown in Figure 1.1. It can be observed in Figure 1.1, a pre-trained model is needed in advance. While two pre-trained models are needed in the proposed approach. Designing of the pre-trained models are described in Section 5.1.1. As it can be observed from Figure 1.1, the ABE process consists of four main processes: estimation of the high-band features, NB features extraction, bandwidth extension process, and narrowband signal reconstruction process. Moreover, the proposed ABE approach utilizes two additional processes to adjust the energy level of the synthesized high-band signal. In this work, the (encoded) narrowband signal is enhanced by the proposed bandwidth extension approach. Section 5.1.2 has an explanation of the proposed artificial bandwidth extension process used at the receiver end.

5.1.1 Designing of the pre-trained models

This section contains the designing processes of two pre-trained models. The designing processes of the pre-trained models consist of the features extraction process and the DNN model designing process, as shown in Figure 5.1 [70]. The features extraction process includes the

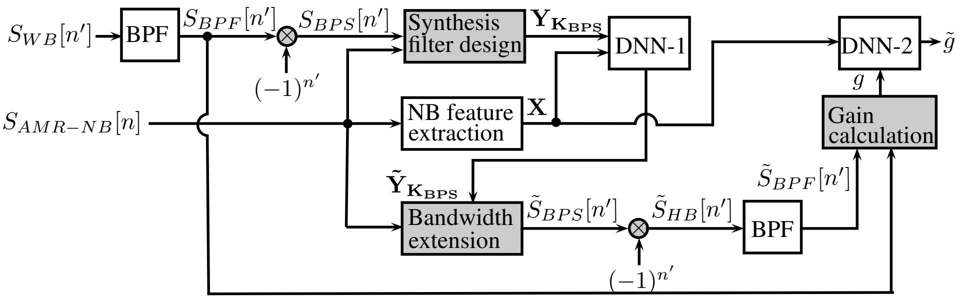

Figure 5.1: Illustrating the training of Deep Neural Networks.

derivation of three features: band pass shifted feature vector $\mathbf{Y_{K_{BPS}}}$, narrowband feature vector $\mathbf{X}$, and gain factor g. For computing these features, two input signals, the (encoded) narrowband signal $S_{AMR-NB}[n]$ and the wideband signal $S_{WB}[n']$ are needed in advance. Besides these input signals, one intermediate band pass shifted signal $S_{BPF}[n']$ is needed. These signals are obtained by following the processes described in Section 5.1.1.1. Processes for computing the features $\mathbf{Y_{K_{BPS}}}$, $\mathbf{X}$, and g are described in Section 5.1.1.2, Section 5.1.1.3, and Section 5.1.1.5, respectively. In Figure 5.1, two DNN models are trained to design the pre-trained models. Separate DNN models DNN-1 and DNN-2 are designed for modeling the band pass shifted feature vector $\mathbf{Y_{K_{BPS}}}$ and gain factor g, respectively. In addition, the gain factor is computed and modeled in such a way that the gain factor reduces the performance loss obtained due to error in the predicted synthesis filter. A description for designing the models DNN-1 and DNN-2 is given in Section 5.1.1.4 and Section 5.1.1.6, respectively.

5.1.1.1 Speech file operations

Speech files are processed for generating speech signals as per the ITU-T protocols at the transmitter side [2, 56]. Speech files, recorded at 16000 Hz sampling frequency and 16 bits per sample, are processed to produce the narrowband signal encoded at 12.2 kbps and reference wideband signal for realistic telephone speech [2, 56, 89]. In addition, the reference band pass shifted signal or mapped high-band signal is needed in the proposed approach, which is generated using the reference wideband signal. Processes of producing these signals are explained in this section.

Narrowband signal production process

A process is drawn in Figure 5.2, which produces an adaptive multi-rate (AMR) coded narrowband signal. In Figure 5.2, the speech signal sampled at 16 kHz is passed through the standard mobile station input (MSIN) high pass filter [2] and then scaled to an active speech level of -26 dBov according to the ITU-T P.56 [90]. The resulting speech signal is filtered by the standard high-quality low pass filter (HQ2) [2] for removing high-frequency components, which gives a narrowband signal $S_{HQ2-MSIN}[n']$ sampled at 16 kHz. n' denotes the sample index for 16000 Hz sampling frequency. The signal $S_{HQ2-MSIN}[n']$ is downsampled by a factor of 2 using the downsampler. Thus, the obtained narrowband signal sampled at 8 kHz is gone through an AMR block, which produces the AMR coded narrowband signal $S_{AMR-NB}[n]$, as shown in Figure 5.2. n represents the sample index for 8000 Hz sampling frequency. The

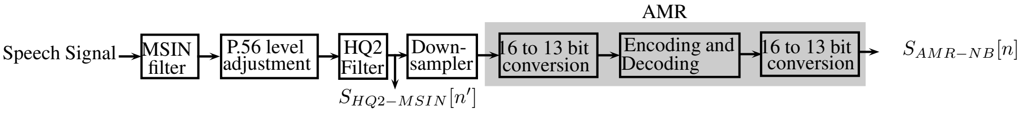

Figure 5.2: AMR coded narrowband signal production process.

AMR block consists of 16 to 13 bit conversion process, encoding using the adaptive multi-rate (AMR) narrowband speech codec at 12.2 kbps and decoding process [89], and again 16 to 13 bit conversion process. The signal $S_{AMR-NB}[n]$ is further enhanced using the proposed ABE framework for synthesizing the frequency components up to 7 kHz.

Wideband signal production process

The speech signal sampled at 16 kHz is passed through the P.341 filter [2] and then scaled to an active speech level of -26 dBov, which leads to an output signal $S_{WB}[n']$, as shown in Figure 5.3. The signal $S_{WB}[n']$ is taken as the reference wideband signal. The reference wideband signal is further used for obtaining the reference band pass shifted signal and for performance analysis.

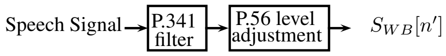

Figure 5.3: Wideband signal production process.

Band pass shifted signal (mapped high-band signal) production process

The reference wideband signal is filtered by a band pass filter (BPF), as shown in Figure 5.4. The resulting signal is taken as a reference band pass filtered signal $S_{BPF}[n']$. Here, the band pass filter passes the frequency components between 4000-7000 Hz (approximately). This band pass filter is designed by the least square method with the specifications: 40 filter order, lower stopband frequency of 3660 Hz (called as stopband frequency1), lower passband frequency of 4340 Hz (called as passband frequency1), higher passband frequency of 7000 Hz (called as passband frequency2), and higher stopband frequency of 7800 Hz (called as stopband frequency2). The BPF is designed using a MATLAB (2019) command *designfilt* and subsequently multiplied by the Kaiser window with a shape factor of five. The signal $S_{BPF}[n']$ has information in the range of 4000-7000 Hz. The high-frequency components of the signal $S_{BPF}[n']$ are shifted into the narrowband region by modulating the signal $S_{BPF}[n']$ with $(-1)^{n'}$, which yields the reference band pass shifted signal (mapped high-band signal) $S_{BPS}[n']$. Similarly, a reverse procedure is applied while synthesizing the band pass filtered speech signal.

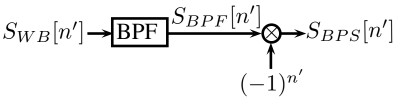

Figure 5.4: Band pass shifted signal production process.

All these speech file operations are conducted for each 20 ms frame duration. Each frame is multiplied with the Hanning window's square root, keeping 50% overlap between the adjacent frames in the proposed artificial bandwidth extension framework. The estimated wideband signals (wideband frames) are multiplied with the Hanning window's square root and subsequently combined using the overlap-add method while reconstructing the whole speech signal [71, 72].

5.1.1.2 Band pass shifted feature vector extraction

The band pass shifted feature vector has information of the proposed synthesis filter. The synthesis filter is used in the bandwidth extension process of the (encoded) narrowband signal. The synthesis filter has high-band envelope information of a signal, which is present in the narrowband region of the synthesis filter. The synthesis filter is designed by using the H^∞-optimization. A system is proposed for designing the synthesis filter. The system is built by combining the process of producing the coded narrowband signal from the narrowband signal $S_{HQ2-MSIN}[n']$ (see Figure 5.2), bandwidth extension process (see Figure 5.1) employed at the receiver side, and reference band pass shifted signal $S_{BPS}[n']$ (see Figure 5.4). This system is drawn in Figure 5.5. The output of this system is an error $e[n']$ between the reference band pass shifted signal $S_{BPS}[n']$ and estimated band pass shifted signal $\tilde{S}_{BPS}[n']$. The estimated band pass shifted signal is an output of the bandwidth extension process. The bandwidth extension process is applied to the coded narrowband signal $S_{AMR-NB}[n]$ (see Figure 5.1 and Figure 5.5). ↓ 2 depicts the downsampler with a downsampling factor of 2.

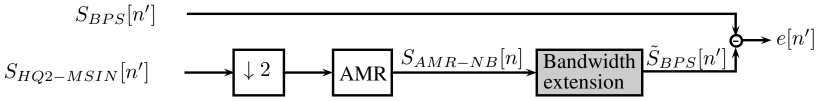

Figure 5.5: A proposed error system.

The error system has two inputs $S_{HQ2-MSIN}[n']$, $S_{BPS}[n']$, and one output $e[n']$. The two inputs of the error system can be converted into a single input by considering the input signal's model. The signal model consists of the spectral envelope information of a signal. Further, the signals $S_{HQ2-MSIN}[n']$ and $S_{BPS}[n']$ are represented by their respective signal models. These signal models are included in Figure 5.5. Therefore, a modified error system is drawn in Figure 5.6.

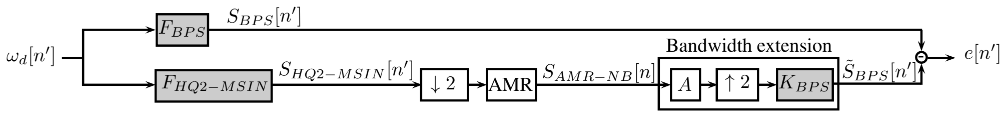

Figure 5.6: A proposed error system considers the signal modeling.

In Figure 5.6, F_{BPS} and $F_{HQ2-MSIN}$ are signal models of $S_{HQ2-MSIN}[n']$ and $S_{BPS}[n']$ signals, respectively. The signals $S_{HQ2-MSIN}[n']$ and $S_{BPS}[n']$ are generated using the excitation signal $\omega_d[n']$ with known features (with finite energy, specifically $\omega_d \in \ell^2(\mathbb{Z}, \mathbb{R}^n)$). The signal models are designed by the Matlab function *prony* based on Prony's method [91]. This function takes three input parameters. The first input parameter is an impulse response. The impulse response is the signal itself in our case. The other two parameters are the number of zeros and poles. The number of zeros and poles are empirically chosen 1, 15 for designing $F_{HQ2-MSIN}$, respectively, and 3, 15 for designing F_{BPS}. The *prony* function returns the numerator and denominator coefficients for the transfer function of a signal model. A few poles and zeros of the signal model may lie outside of the unit circle. However, a minimum phase system is used in the H^∞ optimization problem. Therefore, those poles and zeros of the signal model lying outside the unit circle are reflected inside the unit circle. It can be done by inverting their magnitudes to get the minimum phase system [40]. As a result, the magnitude spectrum of the signal model is not affected; however, the phase spectrum is changed. This will not affect the ABE system as the human auditory system is less sensitive to phase information [40]. The signal models F_{BPS} and $F_{HQ2-MSIN}$ in Figure 5.6 denote the signal models G_1 and G_2 defined in (1.3), respectively. The signal models G_1 and G_2 have the spectral envelope information of the band pass shifted signal (16 kHz) and the narrowband signal (16 kHz), respectively. In this chapter, the signal of interest $S_{BPS}[n']$ has original high-band information in the narrowband region.

In Figure 5.6, the bandwidth extension process is given. This process consists of the LP analysis filter A, upsampler with an upsampling factor, and synthesis filter K_{BPS}. For computing filter A, an all-pole model (order 11) of the signal $S_{AMR-NB}[n]$ is obtained using the linear prediction (LP) analysis [5]. Further, filter A is obtained by inverting the all-pole model. The signal $S_{AMR-NB}[n]$ is fed to the analysis filter A. The output of filter A is a narrowband residual signal. The narrowband residual signal is upsampled by a factor of 2 and subsequently filtered by the synthesis filter K_{BPS}.

Problem formulation

The filter K_{BPS} is designed by following optimization problem.

Problem 4. Given the signal models $F_{HQ2-MSIN}$, F_{BPS}, and analysis filter A, design an optimal stable and causal filter $K_{BPS_{opt}}$ defined as

$$K_{BPS_{opt}} := \arg\min_{K_{BPS}}(\|\mathbb{T}\|_\infty), \tag{5.1}$$

where $\mathbb{T}$ is the discrete error system defined as

$$\mathbb{T} := F_{BPS} - K_{BPS}(\uparrow 2)\ A\ (\text{AMR})(\downarrow 2)F_{HQ2-MSIN}, \tag{5.2}$$

with input $\omega_d[n']$ and output $e[n']$ (see Figure 5.6). Here, $\|\mathbb{T}\|_\infty$ represents the H^∞-norm of the system $\mathbb{T}$.

Solution of Problem 4

Problem 4 is solved for designing the filter K_{BPS} used in the bandwidth extension process. To make Problem 4 mathematically tractable, an ideal AMR block (i.e., AMR=1) is assumed only for solving Problem 4.

The error system $\mathbb{T}$ is converted into the generalized error system (see Figure B.1) as follows

$$\begin{aligned} G_1(z) &= F_{BPS}(z), \\ G_2(z) &= F_{HQ2-MSIN}(z), \\ G_3(z) &= A(z), \\ K_d(z) &= K_{BPS}(z). \end{aligned} \tag{5.3}$$

Further, Problem 4 is solved using the solution given for the generalized error system in Appendix B. The obtained synthesis filter K_{BPS} consists of the high-band spectral envelope information of a signal in the narrowband region. An impulse response of the filter K_{BPS} has infinite terms, i.e., the filter K_{BPS} is an infinite impulse response (IIR) filter. It needs to be converted into a finite impulse response (FIR) for taking it in practical usage. This is done by truncating the Taylor series. The number of terms in the FIR synthesis filter is chosen 15 empirically (see Section 5.2.3.1). The FIR synthesis filter is taken as the band pass shifted feature vector $\mathbf{Y}_{\mathbf{K_{BPS}}}$.

5.1.1.3 Narrowband feature vector extraction

The narrowband envelope information is taken in terms of the sixteen line spectral frequencies (LSFs) [16]. The LSFs are computed for the coded narrowband signal $S_{AMR-NB}[n]$. Also, five other features such as kurtosis, zero-crossing rate, spectral centroid, gradient index, and normalized relative frame energy are taken for capturing the detailed attributes of signal $S_{AMR-NB}[n]$ [13, 92, 93]. These five features and the LSFs are concatenated. The resulting feature vector is represented by $\boldsymbol{x}_i \in \mathbb{R}^{21}$. Further, temporal characteristics are taken into account by considering adjacent frame's information. The final narrowband feature vector of 63 dimensions is constructed similarly to [13]. The narrowband feature vector $\mathbf{X}$ is composed as

$$\mathbf{X} = \left[\boldsymbol{x}_i, \quad \boldsymbol{x}_{i+1} - \boldsymbol{x}_{i-1}, \quad \boldsymbol{x}_{i+1} - 2\boldsymbol{x}_i + \boldsymbol{x}_{i-1} \right],$$

where i, $i-1$, and $i+1$ denote present frame, previous frame, and next frame, respectively.

5.1.1.4 Designing of DNN-1 model

The DNN-1 model is designed using the narrowband feature vector $\mathbf{X}$ and band pass shifted feature vector $\mathbf{Y_{K_{BPS}}}$. The feature vector $\mathbf{X}$ is taken as the input of the DNN-1 model. The feature vector $\mathbf{Y_{K_{BPS}}}$ is taken as the output of the DNN-1 model. The min-max normalization has been applied to the input vector of the DNN-1 model. No normalization is applied to the output of the DNN-1 model. Mean squared error as a loss function is selected for training the DNN-1 model (see DNN-R in [13]).

5.1.1.5 Gain factor computation

The gain factor is computed for adjusting the energy of the estimated band pass filtered signal. The narrowband feature vector $\mathbf{X}$ is fed to the DNN-1 model, which yields an estimated band pass shifted feature vector $\tilde{\mathbf{Y}}_{\mathbf{K_{BPS}}}$. The estimated band pass shifted feature vector $\tilde{\mathbf{Y}}_{\mathbf{K_{BPS}}}$ is used in the bandwidth extension process for estimating the high-band spectral envelope, as shown in Figures 5.1, 5.6, and 5.7. The resulting signal is the estimated band pass shifted signal $\tilde{S}_{BPS}[n']$, which has high-band envelope information in the narrowband region. Therefore, the narrowband region of the signal $\tilde{S}_{BPS}[n']$ is shifted into the high-band region by modulating the signal $\tilde{S}_{BPS}[n']$ with $(-1)^{n'}$. The modulated signal is an estimated high-band signal $\tilde{S}_{HB}[n']$. Further, the signal $\tilde{S}_{HB}[n']$ passes through the band pass filter, which yields the estimated band pass filtered signal $\tilde{S}_{BPF}[n']$. The gain factor is computed as follows

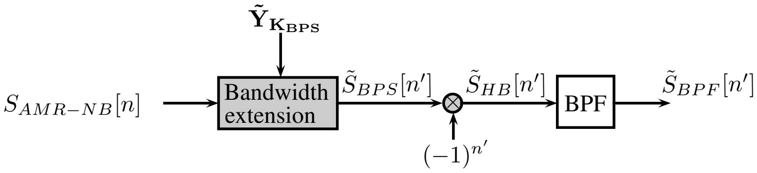

Figure 5.7: Estimation of the band pass filtered signal.

$$g = \sqrt{\frac{\sum_{n'=1}^{N} S_{BPF}^2[n']}{\sum_{n'=1}^{N} \tilde{S}_{BPF}^2[n']}}, \tag{5.4}$$

where $S_{BPF}[n']$ and $\tilde{S}_{BPF}[n']$ are the reference band pass filtered signal and the estimated band pass filtered signal, respectively.

5.1.1.6 Designing of DNN-2 model

The DNN-2 model is designed using the narrowband feature vector $\mathbf{X}$ and gain factor g. The feature vector $\mathbf{X}$ is taken as the input of the DNN-2 model. $2\log_{10} g$ is taken as the output of the DNN-2 model. The min-max normalization is applied to the input vector of the DNN-2 model. The mean and variance normalization is applied to the output of DNN-2 model. Mean squared error as a loss function is selected for training the DNN-2 model (see DNN-R in [13]).

5.1.2 Extension of the AMR coded narrowband signal

This section has a discussion of the artificial bandwidth extension process, as outlined in Figure 5.8. Pre-trained models DNN-1 and DNN-2 are used in the artificial bandwidth extension process. The artificial bandwidth extension process involves five main processes: narrowband signal reconstruction, band pass shifted feature vector prediction, gain factor prediction, high-band signal estimation, and wideband signal synthesis, as elaborated in Sections 5.1.2.1, 5.1.2.2, 5.1.2.3, 5.1.2.4, and 5.1.2.5, respectively.

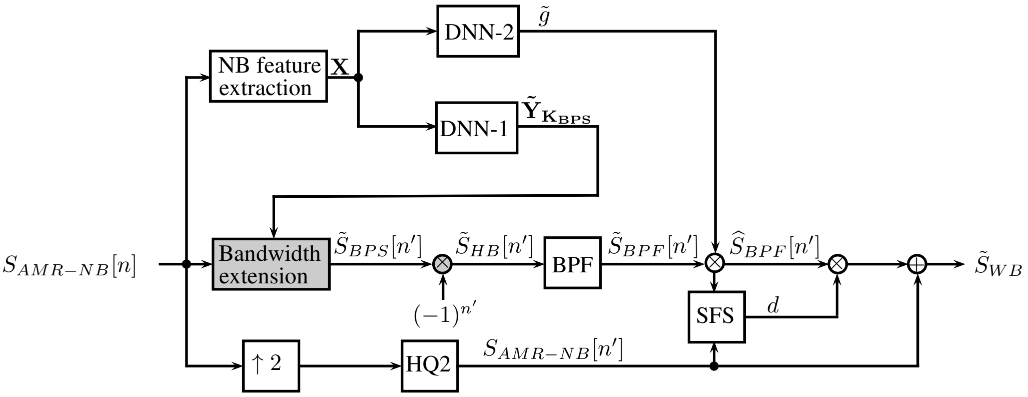

Figure 5.8: Illustrating the artificial bandwidth extension of the coded narrowband signal.

5.1.2.1 Narrowband signal reconstruction process

The narrowband signal reconstruction process is used to resample the narrowband signal. The AMR coded narrowband signal $S_{AMR-NB}[n]$ sampled at 8 kHz is resampled at 16 kHz. For this, the signal $S_{AMR-NB}[n]$ is upsampled by a factor of 2 and subsequently filtered by the

HQ2 low pass filter. This leads to an output signal $S_{AMR-NB}s[n']$ sampled at 16 kHz, as shown in Figure 5.8.

5.1.2.2 Band pass shifted feature vector prediction

The band pass shifted feature vector is estimated, which is used in bandwidth extension of the AMR coded narrowband $S_{AMR-NB}[n]$. For this, the narrowband feature vector $\mathbf{X}$ is computed using the signal $S_{AMR-NB}[n]$. The feature vector $\mathbf{X}$ is normalized by min-max normalization. Then, the normalized narrowband feature vector $\mathbf{X}$ is fed to the DNN-1 model, which produces the estimated band pass shifted feature vector $\tilde{Y}_{K_{BPS}}$.

5.1.2.3 Gain factor prediction

The gain factor is predicted for adjusting energy level of the estimated band pass filtered signal. For this, the min-max normalized feature $\mathbf{X}$ as computed in Section 5.1.2.2 is fed to the DNN-2 model. Further, an output of the DNN-2 model is de-normalized by applying the reverse mean and variance normalization procedure, which yields a scalar value $\tilde{g}_1$. Further, the estimated gain factor $\tilde{g}$ is computed as $\tilde{g} = 10^{(\tilde{g}_1/2)}$. .

5.1.2.4 High-band signal estimation

The high-band signal $\tilde{S}_{HB}[n']$ is estimated using the predicted band pass shifted feature vector $\tilde{\mathbf{Y}}_{\mathbf{K_{BPS}}}$. The signal $\tilde{S}_{HB}[n']$ for a given signal $S_{AMR-NB}[n]$ is obtained by following Figures 5.6 and 5.7.

5.1.2.5 Wideband signal synthesis

The wideband signal is estimated by adding the resampled narrowband signal and the modified estimated high-band signal obtained using the estimated gain factor and an attenuation factor. The estimated gain factor and attenuation factor are used to control the energy level of the estimated high-band signal. For this, the signal $\tilde{S}_{HB}[n']$ is fed into the band pass filter for extracting the desired frequency components, as shown in Figure 5.7. The obtained signal $\tilde{S}_{BPF}[n']$ is multiplied with the estimated gain factor $\tilde{g}$, which gives an output signal $\widehat{S}_{BPF}[n']$.

The spectral floor suppression (SFS) technique [13] is used for controlling the synthesized energy of the signal $\widehat{S}_{BPF}[n']$. Further, a ratio R^{SFS} is computed as

$$R^{\text{SFS}} = \frac{\frac{1}{N/4}\sum_{k=(N/4)+2}^{(N/2)+1} \left|\widehat{\phi}_{BPF}[k]\right|^2}{\frac{1}{(N/4)+1}\sum_{k=1}^{(N/4)+1} \left|\widehat{\phi}_{AMR-NB}[k]\right|^2}, \tag{5.5}$$

where $\widehat{\phi}_{BPF}[k]$ and $\widehat{\phi}_{AMR-NB}[k]$ are power spectrum densities of the signals $\widehat{S}_{BPF}[n']$ and $S_{AMR-NB}[n']$, respectively. An attenuation factor is calculated as

$$d = \min\left\{\frac{d_{\text{high}} - d_{\text{low}}}{\theta^{\text{SFS}}} R^{\text{SFS}} + d_{\text{low}}, d_{\text{high}}\right\} \text{dB}, \tag{5.6}$$

where $d_{\text{high}} = -7$ dB, $d_{\text{low}} = -13$ dB, and $\theta^{\text{SFS}} = 5$. These parameter values have been chosen empirically in the proposed ABE framework.

Finally, the wideband signal is estimated by adding the signal $S_{AMR-NB}[n']$ and an estimated high-band signal obtained by applying the attenuation factor d on $\widehat{S}_{BPF}[n']$ defined as

$$\tilde{S}_{WB}[n'] = S_{AMR-NB}[n'] + 10^{\frac{d}{20}}\widehat{S}_{BPF}[n'], \tag{5.7}$$

where $\tilde{S}_{WB}[n']$ is the estimated wideband signal.

5.2 Speech databases, measures, and results analysis

Experiments are conducted to analyze the performance of the proposed ABE framework. The Performance of the proposed ABE approach is analyzed on speech samples, which are described in Section 5.2.1. Measures are chosen for analyzing the performance of the proposed ABE framework, which are discussed in Section 5.2.2. Results are discussed in Section 5.2.3.

5.2.1 Databases

The proposed ABE framework is analyzed on the two datasets: TIMIT dataset [73] and RSR15 dataset [74]. These datasets have speech samples, which are recorded at 16000 Hz sampling frequency and 16 bits per sample. The speech samples are processed, as done in Section 5.1.1.1. The TIMIT dataset is already segmented into two sets: training set and test set. The training set is used for training the DNN models, while the test set is considered as a validation set in our analysis. A test set is constructed by taking speech files from the RSR15 dataset. The test set consists of the speech files spoken by 4 female and 3 male speakers. The test set is taken from a different database, which leads to more generalized results.

5.2.2 Measures for performance evaluation

For measuring the performance, we use the wideband perceptual evaluation of speech quality (PESQ) in terms of the wideband mean opinion score listening quality objective (MOS-LQO) [81,82], upper-band (4-7 kHz) logarithmic spectral distance (LSD$_{UB}$), and full-band (0-7 kHz) logarithmic spectral distance (LSD$_{FB}$) (see Appendix C.) [78].

5.2.3 Results analysis

In this section, results are analyzed and discussed. The IIR synthesis filter K_{BPS} is to be converted into an FIR synthesis filter to take it in practical usage. The number of terms in the FIR synthesis filter is decided in such a way that the FIR synthesis filter gives the best wideband MOS-LQO using a DNN architecture. It is started with choosing the number of terms 15 in the FIR synthesis filter. Initially, the DNN architecture is designed for the FIR synthesis filter length 15 and then compared with the other lengths such as 5, 10, 20, and 25. The coefficients of each FIR synthesis filter are divided by the maximum value of coefficients before designing the deep neural network architecture. The DNN model is then designed in Section 5.2.3.1 to decide the synthesis filter length and model the synthesis filter. Another DNN model is designed for modeling the gain factor in Section 5.2.3.2. The proposed approach is compared with two baselines in Sections 5.2.3.3 and 5.2.3.4. Section 5.2.3.3 has the objective comparison. Section 5.2.3.4 has the subjective comparison.

5.2.3.1 DNN-1 model architecture

An architecture of DNN is decided using the feature vectors $\mathbf{X}$ and $\mathbf{Y_{K_{BPS}}}$ empirically. The feature vector $\mathbf{Y_{K_{BPS}}}$ consists of the coefficients of the FIR synthesis filter. The feature vectors $\mathbf{X}$ of all the sets (training set, validation set, and test set) are normalized using the statistics, which are computed using the training set only. The batch size and learning rate have been fixed to 128 and 0.00001 for training the DNN model, respectively. The activation function ReLU has been set for the hidden layers, and linear has been set for the output layer. L2 and L1 regularization for the output layer weights are employed to avoid the risk of over-fitting [70]. Values of L2 and L1 regularizations are fixed to 0.0001 and 0.0001, respectively. A stopping criterion is chosen as the minimum validation error. Training of the DNN model is stopped if the validation error does not improve for 7 epochs. Different DNN architectures are trained and designed by varying the number of hidden-layers and the number of hidden-layer neurons. Predicted outputs of the validation set, generated from different DNN architectures, are used in the bandwidth extension approach. Here, the bandwidth extension approach is implemented without the SFS technique, i.e., $d = 0$ dB. Also, the gain factor g corresponding to the predicted synthesis filter, used in bandwidth extension, is computed using (5.4). The wideband MOS-LQO values are computed for the extended speech signals of the validation set using different DNN architectures and then compared in Table 5.1.

Table 5.1: Computation of wideband MOS-LQO for the validation set with varying the DNN architecture

Number of hidden-layers	2	2	2	2	3	3	3	3
Number of Neurons in each hidden-layer	32	64	128	256	32	64	128	256
Wideband MOS-LQO	3.4024	3.4216	3.3846	3.3804	3.3716	3.3717	3.4041	3.4073

This DNN architecture with 2 hidden layers and 64 neurons in each hidden layer is decided, which gives the best wideband MOS-LQO value for the validation set. Further, this architecture is trained for the other lengths such as 5, 10, 20, and 25. The wideband MOS-LQO for the validation set is computed by varying the FIR synthesis filter length used in the DNN model and listed in Table 5.2. The wideband MOS-LQO value is obtained better by using the filter

Table 5.2: Wideband MOS-LQO computation for the extended speech signals of the validation set using the DNN architecture designed with 2 hidden layers and 64 neurons.

FIR synthesis Filter Length	5	10	15	20	25
MOS-LQO	3.3377	3.3646	3.4216	3.3852	3.3620

length 15 than the other lengths. DNN architecture with 2 hidden layers and 64 neurons in each hidden layer is further named as the DNN-1 model used for estimating the FIR synthesis filter.

Furthermore, parameters d_{high}, d_{low}, and θ^{SFS} used in the spectral floor suppression are decided empirically. These parameters are decided based on the wideband MOS-LQO value for the validation set. It is done using the predicted outputs from the DNN-1 model in the bandwidth extension process and computing the gain factor using (5.4). The values of d_{high}, d_{low}, and θ^{SFS} are chosen -7 dB, -13 dB, and 5 over a wide range, respectively. These parameters value is chosen in such a way that the SFS technique reduces noise artifacts present in speech sounds. The MOS-LQO value for the validation set is obtained 3.7502 points using these values.

5.2.3.2 DNN-2 model architecture

Another architecture of DNN is designed for estimating the gain factor, which is designed by using the narrowband feature vector $\mathbf{X}$ and gain factor g. $\mathbf{X}$ and g are normalized using the statistics obtained for the training set. $\mathbf{X}$ is normalized using the min-max normalization, while g is normalized using mean and variance normalization. The batch size and learning rate are chosen 512 and 0.001, respectively. The L2 regularization for the layer weights has been used. The value of the L2 regularization is chosen 0.00001. The stopping criterion is selected as the minimum validation error. Different DNN architectures, made by varying the number of hidden layers and the number of neurons, are then trained. These architectures are tested on the validation set, as done in designing the DNN-1 model. The wideband MOS-LQO values are computed for the extended speech signals of the validation set using different DNN architectures and then compared in Table 5.3.

Table 5.3: Computation of wideband MOS-LQO for the validation set with varying the DNN architecture

Number of hidden-layers	2	2	2	2	3	3	3	3
Number of Neurons in each hidden-layer	128	256	512	1024	128	256	512	1024
Wideband MOS-LQO	2.9362	2.9361	2.9776	2.9196	2.9476	2.9775	2.9644	2.9241

Table 5.4: Performance evaluation on the test set for the proposed approach and baselines.

Method	LSD$_{UB}$ (dB)	LSD$_{FB}$ (dB)	MOS-LQO
Proposed approach	15.8544	12.1325	3.3400
Modulation technique	19.8696	14.5247	3.1332
Cepstral domain	11.4070	9.7390	2.7540

An architecture designed with 2 hidden layers and 512 neurons in each hidden layer is selected, which produces the best wideband MOS-LQO for the validation set. This architecture is selected as the DNN-2 model.

5.2.3.3 Objective comparison with baselines

In this section, a comparison of the proposed ABE framework with two baselines is discussed. The baselines are the cepstral domain approach [39] and the modulation technique [13]. The cepstral domain ABE approach synthesizes the high-band information using the high-band magnitude spectrum of a signal. The high-band magnitude spectrum is obtained by the linear frequency cepstral coefficients, which are predicted using a DNN model. The phase of high-band region is directly estimated using the phase of narrowband region. The modulation technique is based on the source-filter model. The high-band envelope information in the modulation technique is obtained by the linear frequency cepstral coefficients, which are predicted using a DNN model. For estimating the high-band residual, the spectral translation method is utilized. Experimental conditions such as window duration, type of window, datasets, and narrowband processing have been fixed in our implementation of the baselines and proposed approach.

The objective measures are computed for the proposed framework and baselines on the test set as listed in Table 5.4. As it can be observed from Table 5.4, the proposed method improves the MOS-LQO value by 0.2068 and 0.5860 points compared to the modulation technique and cepstral domain approach, respectively. The proposed method reduces the upper-band logarithmic spectral distance (LSD$_{UB}$) by 4.0152 dB and the full-band logarithmic spectral distance (LSD$_{FB}$) by 2.3922 dB when compared to the modulation technique. The proposed method increases the LSD$_{UB}$ by 4.4474 dB and the LSD$_{FB}$ by 2.3935 dB when compared to the cepstral domain approach. The proposed method produces the moderate LSD$_{UB}$, LSD$_{FB}$, and the best MOS-LQO when compared to the baselines. The better MOS-LQO may produce better speech quality of the speech signal. The cepstral domain approach produces the best LSD$_{UB}$, LSD$_{FB}$, and the worst MOS-LQO. The worst MOS-LQO may give more noise artifacts in the extended speech signal. The modulation technique gives the worst LSD$_{UB}$, LSD$_{FB}$, and the moderate MOS-LQO. The worst LSD may result less perception of speech sounds.

The spectrogram of a speech file is further observed and discussed. For this, a female speech file is taken from the test set. The spectrogram of the female speech file is shown in Figure 5.9. Figure 5.9 **(a)**, **(b)**, **(c)**, **(d)**, and **(e)** show spectrogram of the reference female speech signal, AMR coded narrowband speech signal sampled at 16 kHz, bandwidth extended speech signals by the proposed approach, modulation technique, and cepstral domain approach, respectively. It can be observed in Figure 5.9 **(e)**, a pattern like noise is seen in spectrogram of the extended speech signal by the cepstral domain approach. As a result, energy in the estimated high-band region is high. But this noise affects the speech quality. However, this noise is not seen in Figure 5.9 **(c, d)**. Therefore, the speech quality is better for the proposed approach and modulation technique than the cepstral domain approach. Energy present in the high-band region of bandwidth extended speech signal is higher for the proposed approach than the modulation technique. This may produce a better perception of speech sounds for the proposed approach than the modulation technique. Some noise may be present in the extended speech signal generated by the proposed approach, however, it does not affect the perception of sounds.

5.2.3.4 Subjective comparison

A subjective listening test is conducted to examine the perceptual quality of speech signals. It has been done by following ITU-T P.800 [86, Annex E]. In the listening test, two speech files are compared to each other by considering the speech characteristics like noise artifacts, perception, sound level, and overall speech quality. Rating is given on a scale from -3 (much worse) to 3 (much better). The rating scale is named the CMOS (comparison mean opinion

Table 5.5: Subjective assessment conducted on the artificially extended speech files belonging to the test set.

Conditions	CMOS	CI_{95}
Re-sampled AMR coded narrowband signal ($S_{AMR-NB}[n']$) vs Proposed approach	1.3944	[1.2300 1.5589]
Modulation technique vs Proposed approach	1.1833	[1.0182 1.3485]
Cepstral domain approach vs Proposed approach	1.2556	[1.0681 1.4430]

score) scale. Fifteen listeners have participated in the subjective assessment. These listeners do not have any hearing problems. Their ages are between 25 to 32 years. CMOS score is computed for the three conditions. The first condition is that artificially extended speech files by the proposed approach are compared with their corresponding re-sampled AMR coded narrowband signals. Rest conditions are: the artificially extended speech files by the proposed approach are compared to the artificially extended speech files by the modulation technique and cepstral domain approach. Twelve speech files are taken from the test set for listening. The twelve pairs of speech files are compared and then rated for each condition. The speech files are band pass filtered by the P.341 filter [2] and subsequently scaled to an active speech level of -26 dBov [90]. CMOS and 95% confidence interval are listed in Table 5.5 for each condition. The proposed approach improves the CMOS value by 1.3944, 1.1833, and 1.2556 points when compared to the re-sampled AMR coded narrowband signal ($S_{AMR-NB}[n']$), extended speech signals by the modulation technique and cepstral domain approach, respectively. In the subjective listening test, listeners gave their opinions. According to opinions, perception of speech sounds is obtained higher in the extended speech signals by the proposed approach than the re-sampled AMR coded narrowband signal ($S_{AMR-NB}[n']$) and extended speech signals by the modulation technique. Noise present in extended speech sounds is suppressed higher by the proposed approach than the cepstral domain approach.

5.3 Objective comparison with the previous schemes

In this section, we compare the performances of different types of signal modeling schemes proposed in the all chapters.

Table 5.6: Performance evaluation of each modeling scheme for the speech files belonging to the validation set.

Signal modeling schemes	LSD$_{UB}$ (dB)	LSD$_{FB}$ (dB)	MOS-LQO
Wideband modeling	9.0480	7.8191	3.1904
High-band modeling	9.0159	7.7843	3.5540
Mapped high-band modeling	8.0167	7.4869	3.7962

5.3.1 An objective comparison in the oracle conditions

Our proposed approaches (in chapters 3, 4, and 5) are mainly different in respective to signal modeling scheme or signal of interest. To do a fair comparison, all these approaches are implemented here keeping the same experimental conditions except the signal modeling schemes. Experimental conditions means the encoded narrowband signal at 12.2 kbps, Hanning window, wideband signal, and reconstruction process (refer Chapters 3 and 5). The error system $\mathbb{T}$ is defined for all the modeling schemes as follows

$$\mathbb{T} = \left\{ \begin{array}{ll} F_{WB} - K(\uparrow 2)\ A\ (\text{AMR})(\downarrow 2) F_{HQ2-MSIN}, & \text{for wideband modeling} \\ F_{HB} - K(\uparrow 2)\ A\ (\text{AMR})(\downarrow 2) F_{HQ2-MSIN}, & \text{for high-band modeling} \\ F_{BPS} - K(\uparrow 2)\ A\ (\text{AMR})(\downarrow 2) F_{HQ2-MSIN}, & \text{for mapped high-band modeling} \end{array} \right\} \tag{5.8}$$

where F_{WB}, F_{HB}, and F_{BPS} are the signal models of wideband (50-7000 Hz) signal, bandpass (4000-7000 Hz) signal, and bandpass (4000-7000 Hz) shifted signal. Further, high-band speech signals are obtained using the oracle IIR synthesis filters (without using any machine learning modeling technique, i.e., oracle condition) for each modeling scheme. One more thing is that we are not applying any post-processing steps, viz. SFS technique and DFT concatenation to know the effect of signal modeling schemes. Performance of each modeling scheme is presented in Table 5.6 on the speech files belonging to the validation set. It is observed in Table 5.6 that high-band modeling yields more improvement in MOS-LQO than LSD (upper-band and full-band LSD) when compared with wideband modeling, and the mapped high-band modeling improves the MOS-LQO and LSD (upper-band and full-band LSD) when compared with high-

Table 5.7: Performance comparison between wideband modeling and mapped high-band modeling in practical scenario.

Signal modeling schemes	LSD_{UB} (dB)	LSD_{FB} (dB)	MOS-LQO
Wideband modeling	17.6657	13.2050	3.3022
Mapped high-band modeling	15.8544	12.1325	3.3400

band modeling. The mapped high-band modeling scheme among all the modeling schemes gives the best objective measures, as seen in Table 5.6.

5.3.2 An objective comparison in practical conditions

In this section, we discuss the performance of the proposed approaches in Chapters 3 and 5 in practical conditions (using machine learning modeling techniques). Our proposed approaches are mainly different from each other with respect to signal modeling schemes. Chapter 3 uses wideband modeling, and Chapter 5 uses mapped high-band modeling. The objective measures are listed in Table 5.7 of chapters 3 and 5 on the test set. It can be observed in Table 5.7 that the objective measures are obtained better by using the mapped high-band modeling than the wideband modeling.

5.4 Conclusion

This work proposes to use the modulation process, H^∞ optimization, and DNN modeling for obtaining the synthesis filter. The modulation process is used to shift the high-frequencies into the narrowband region for getting better results using the H^∞ optimization. The H^∞ optimization helps in acquiring the synthesis filter corresponding to a signal model (pole-zero model) and an analysis filter. The synthesis filter has the high-band spectral envelope information of a signal in its narrowband region. The gain adjustment and spectral floor suppression techniques are used for controlling the energy of synthesized high-frequency components. Separate DNN models are designed for estimating the gain factor and synthesis filter. DNN modeling and computation of the gain factor reduce the performance loss, which is obtained due to obtaining error in the predicted synthesis filter. The MOS-LQO objective measure is improved by the proposed approach in comparison to the baselines. Also, in subjective listening test, CMOS value is obtained higher by the proposed approach when compared to the baselines.

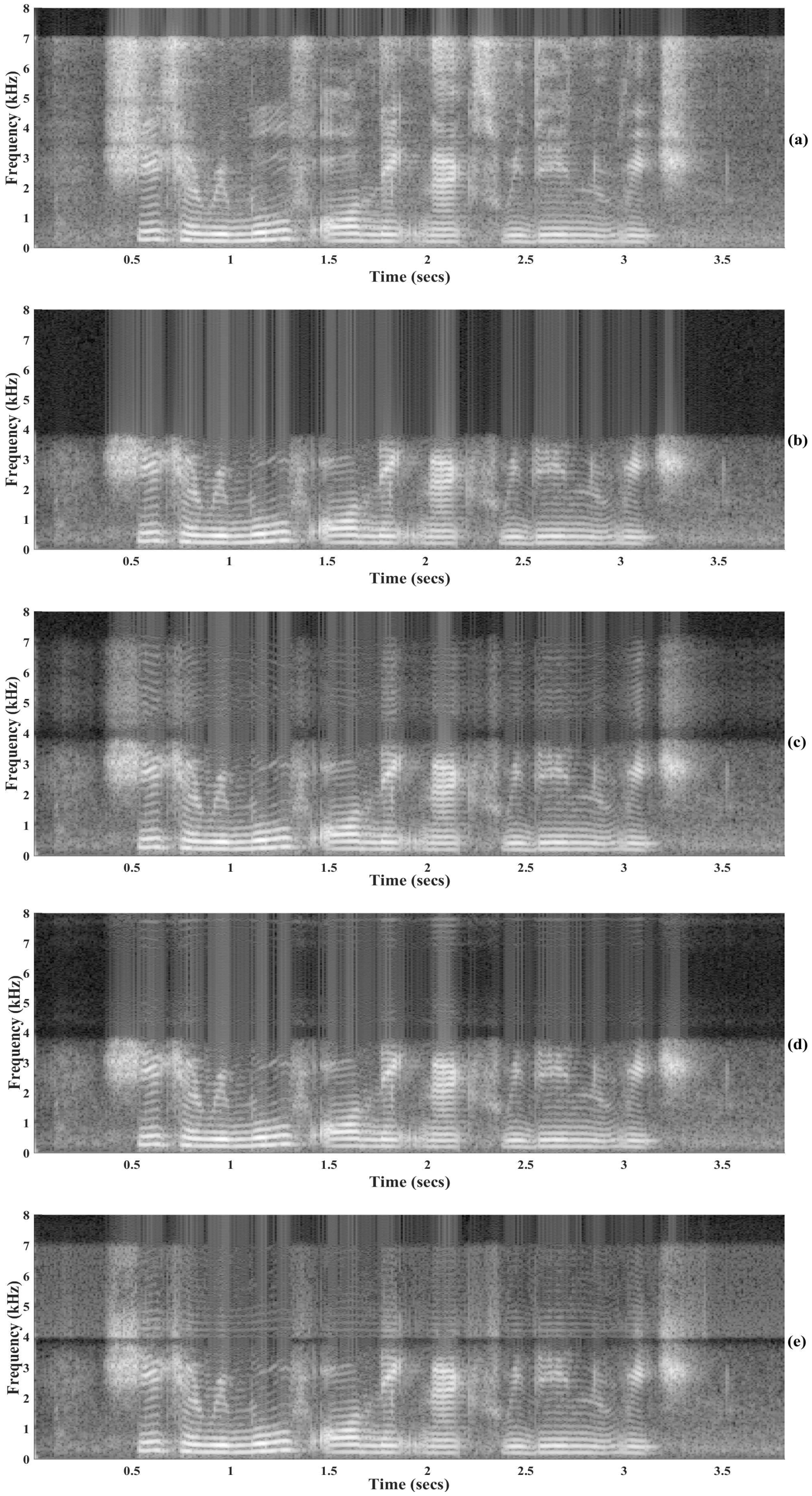

Figure 5.9: Spectrogram of **(a)** reference wideband speech signal of a female speaker, **(b)** AMR coded narrowband signal sampled at 16 kHz, and **(c,d,e)** extended speech signal by the proposed approach, modulation technique, and cepstral domain approach, respectively .

6

Summary and future work

Contents

6.1 Summary of the work 102

6.2 Future directions 104

6.1 Summary of the work

There are two main goals, which are addressed in this thesis. The first goal is to explore the H^{∞} sampled-data control theory in the artificial bandwidth extension process used at the receiver end in communication. As per the theoretical point of view, we can apply this theory to a linear time-invariant system. However, a linear time-invariant system can not be obtained for different speech sounds. Because all speech sounds do not have the same characteristics. Therefore, the H^{∞} sampled-data control theory is alone not sufficient for the speech domain. This theory is applicable only for the short duration of around 10-30 ms, in which we can obtain an LTI system for representing the speech production model. We can design a synthesis filter for the small speech segment using the H^{∞} sampled-data system theory. However, it is not sufficient in the practical scenario. A variety of synthesis filters are designed for different speech segments. It is infeasible to store all the synthesis filters. Therefore, we used machine learning modeling techniques, which store information of synthesis filters in a compact form. The second goal is to use of different types of signal models. The signal model has spectral envelope information of a signal. The spectral envelope of a speech signal consists of poles as well as zeros. In state of art, only poles in the signal model are taken into account. We consider poles as well as zeros in the signal model for better signal modeling. We have experimented with using three types of speech signal models. These signal models depend upon signals spectrum of interest, which are used in designing the synthesis filter. The signal model has spectral envelope information of the signal of interest. In this thesis, we have experimented with considering wideband signal, high-band signal, and mapped high-band signal as the signals spectrum of interest. The mapped high-band signal modeling out of these signal modeling performs best overall. In addition, we enhanced three types of narrowband signals. One is the aliased narrowband signal, the second is the pure narrowband signal, and the third is the encoded narrowband signal (compressed narrowband signal). The major contributions incorporated in this thesis are summarized below:

- Initially, an ABE approach is proposed for the aliased narrowband signal. The aliased narrowband signal has distorted low-frequency components. However, it establishes the better conditional dependency between narrowband and wideband information, which helps in the estimation of the synthesis/interpolation filter. Therefore, GMM and DNN models perform almost the same. In this approach, we estimate the full wideband signal as there is an aliasing distortion in low-frequency components. Therefore, the signal of interest is the wideband signal. As a result, wideband signal modeling is used in the proposed ABE approach. The interpolation filter is obtained using the H^∞ optimization. The obtained interpolation filter is used in the bandwidth extension process of the aliased narrowband signal. This approach is easy to implement but can not be used for the existing transmitter set-ups. However, this approach showed the potential of H^∞ sampled-data system theory in ABE when we focus just on high-frequency signals like unvoiced speech signals.

- We next concentrate upon the standard transmitter set-ups. A new ABE approach is proposed using H^∞ sampled-data system theory, which is compatible with the existing transmitter set-ups. We followed the ITU-T standards as done by peers. It means the band-limited (300-3400 approximately) narrowband signal encoded at 12.2 kbps is enhanced by the proposed ABE approach. The proposed ABE approach also considers wideband signal modeling. The synthesis filter corresponding to the wideband signal model is obtained using H^∞ optimization. This synthesis filter has wideband spectral envelope information. However, the narrowband spectral envelope information in the synthesis filter is not needed, because the narrowband signal is available at the receiver end (due to using the standard Tx set-up). Therefore, the narrowband spectral envelope information is suppressed in the synthesis filter. Further, the gain adjustment and spectral floor suppression techniques are used to control the energy of synthesized high-frequency components. The DNN model is used to estimate the synthesis filter for enhancing an unknown and uncertain speech signal. The DNN model performs better than the GMM model.

- Further, the post-processing applied to the synthesis filter is avoided, as done in the previous approach. Post-processing is included in the optimization problem. For this, we change the signal model. We used the high-band signal modeling in the proposed approach. Also, the proposed ABE approach enhances the narrowband signal consisting of frequency components up to 4 kHz approximately. The proposed approach uses H^{∞} optimization for designing the synthesis filter corresponding to the high-band signal model. The obtained synthesis filter has high-band spectral envelope information. Besides, the gain adjustment technique is used to set the energy level of the estimated high-band signal, and the DFT concatenation is used to avoid the unwanted information leaked by the non-ideal filters (synthesis filter and low pass filter) in the wideband signal estimation. The DNN model is used for predicting the synthesis filter and gain factor.

- Further, we again changed the signal modeling, which leads to better results. We used the mapped high-band signal modeling to get a better solution by the H^{∞} sampled-data system theory. The mapped high-band signal has the high-band information mapped to the narrowband region using modulation. Additionally, we modified the set-ups as per the ITU-T protocols for a better comparison with peers. Apart from that, we use the gain adjustment and spectral floor suppression techniques for controlling the energy of the estimated high-band signal. Separate DNN models are used for estimating the synthesis filter and gain factor. Also, the computation process of the gain factor reduces the performance loss due to obtaining errors in the estimated synthesis filter.

6.2 Future directions

In this thesis, we proposed to use of the H^{∞} sampled-data system theory for artificial bandwidth extension. Our work shows the potential of the H^{∞} sampled-data system theory in speech processing with a lot of possibilities for further research. Therefore, we list some of the possible future directions as follows:

- We obtained the approximated FIR filter of the IIR synthesis filter by truncating the higher-order Taylor series in all the proposed ABE approaches. However, this approach does not provide an optimal selection of coefficients. Therefore, as a future direction, optimal FIR representation of the synthesis IIR filter can be used for all the proposed ABE approaches.

- Signal models are obtained using Prony's method in the proposed ABE approaches. Other signal modeling schemes (such as recursive methods [95, 96], recursive weighted linear least-squares (WLLS) procedure [97], Newton-like algorithm [98], and quasi-Newton algorithm [40]) can be used to see the effect of different signal modeling schemes on the performances.

- A deep study could be done for different phonemes. We can experiment to decide the optimal signal model for each phoneme. It means the optimal number of poles and zeros in the signal model could be decided for each phoneme empirically. It may result that length of the FIR synthesis filter can be different for different phonemes. It needs to design a separate statistical model for each phoneme.

- The band-limited narrowband signal encoded at 12.2 kbps has frequency components between 300-3400 Hz approximately. Low-frequency components in the range of 0-300 Hz can be recovered to improve speech quality. In addition, losses obtained due to encoding of the narrowband signal can be reduced.

- We assumed the ideal AMR block in our work. It may produce error in obtaining the synthesis filter. Results might be better if the exact model of AMR is used.

- An analysis of the performance of the proposed bandwidth approach can be done by using different narrowband signals encoded at different bit rates, such as 4.75 kbps, 5.15 kbps, 5.90 kbps, 6.70 kbps, 7.40 kbps, 7.95 kbps, and 10.20 kbps.

- We could extend the work for the noisy signals.

- The H^∞ sampled-data system theory could be used in other speech applications, such as speaker identification, speaker verification, and speech classification etc.

- The H^∞ sampled-data system theory could be explored for other signals, such as image, audio, ECG, etc.

A

Sampled-data system theory

Contents

A.1 History of sampled-data system theory 108

A.2 Abbreviations 109

A.3 A general closed-loop system 109

A.4 Lifting and inverse lifting 110

A. Sampled-data system theory

In this appendix, we provide a brief introduction to the sampled-data system theory.

A.1 History of sampled-data system theory

It is a well-established result that digital signal transmission has numerous benefits over analog transmission. Therefore, the analog/continuous signal is discretized using the sampling process. The resulting discrete signal is transmitted but reconstructed back to the continuous domain at the receiver. Here the main aim is to recover an analog signal from its samples with minimal error. This is called a (continuous/analog) signal reconstruction problem. The signal reconstruction problem is the fundamental problem in digital signal processing. The sampling theorem ([99]) states that we can recover the original continuous signal from the sampled-data if it is band-limited. In practice, signals are not band-limited. Hence, a popular solution to achieve band-limited signal by using an anti-aliasing filter introduces another distortion due to the Gibbs phenomenon. Furthermore, the impulse response of the ideal anti-aliasing filter is difficult to implement as it is non-causal and does not decay very fast.

To find an optimal answer to the signal reconstruction problem in general, researchers started looking at these problems as mathematical optimization problems (see Sun et al. [100], and Unser [101]). Mathematically, this means the design of an analog to discrete converter (sampler) and a discrete to analog convertor (hold) given the error criterion. The major challenge is in treating discrete and analog signals (or multi-rate) in a common framework. The Sampled data system theory provided such a framework. In 1995, Chen and Francis [48] applied the sampled-data system theory to the signal reconstruction problem entirely in the discrete domain (this problem is at the heart of this thesis). The continuous signal reconstruction problem was studied first in 1996 by Khargonekar and Yamamoto [102]. Instead of aiming at exact reconstruction as in the Shannon case, minimizing the error without throwing away any frequencies is the main criterion in the signal reconstruction using sampled-data system theory. The optimization is done using the H^2-norm or H^∞ criterion. The sampled data system theory is applied to several signal processing applications using different error criteria with or without causality constraints after the Khargonekar and Yamamoto paper [102] in 1996. For example, downsampling with causality constraints (using fast sampler/fast hold approximation) is treated in [103–105], audio compression in [43], image application in [106] etc. For a complete list of applications, see the review paper by Yamamoto et al. [107]. Meinsma and Mirkin [108] applied the sampled-data system theory to the cases where a non-causal sampler (or hold) is fixed, and we have to design hold (or sampler). They have also designed relaxed causal (i.e., with limited access to future) hold given a sampler using sampled data system theory [109,110]. Shekhawat and Meinsma applied sampled data system in non-causal downsampling and in the design of relaxed causal sampler design given a hold [46,111].

In this thesis, we have used the result from Chen-Francis [48], and Yamomoto [107] (which are described next).

A.2 Abbreviations

$\mathbb{Z}$: The set of integers.

LDTI: Linear discrete time invariant.

$\mathbb{R}$: The set of real numbers.

$\mathbb{R}^n$: n-dimensional vector space over $\mathbb{R}$.

$l^2(\mathbb{Z}, \mathbb{R})$: Square summable sequences in $\mathbb{R}^n$.

$||.||_2$: l^2-norm of a discrete sequence.

$\uparrow N$: Upsampler with an upsampling factor N, i.e., inserting the N-1 zero-valued samples between two consecutive original samples for increasing the sampling rate.

$\downarrow N$: Downsampler with a downsampling factor N, i.e., keeping every N^{th} sample and deleting the remaining samples.

A.3 A general closed-loop system

The standard closed-loop system $\mathbb{T}$ is made up of a generalized plant G and a feedback controller K_d, as shown in Figure A.1. The transfer function of system $\mathbb{T}$ from ω to ζ is written

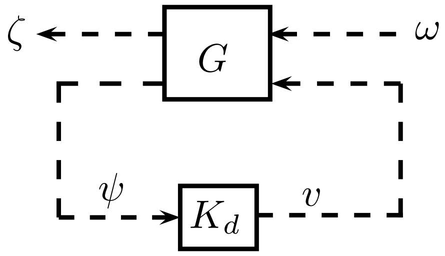

Figure A.1: Single rate discrete-time lifted system [1]

as follows [1]

$$\mathbb{T} = G_{11} + G_{12}K_d(I - G_{22}K_d)^{-1}G_{21}, \tag{A.1}$$

where G is

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}. \tag{A.2}$$

A.4 Lifting and inverse lifting

The lifting technique converts the one-dimensional signal into a multi-dimensional signal and vice versa by inverse lifting [50]. This can be applied for continuous signals and discrete signals. We need only discrete-time lifting and inverse lifting. Discrete-time lifting operator by a factor of N is defined by $\mathbb{L}_N$ in the time domain, and it is defined as [1]

$$\mathbb{L}_N: \quad l^2(\mathbb{Z}, \mathbb{R}) \rightarrow l^2(\mathbb{Z}, \mathbb{R}^N), \tag{A.3}$$

$$\left\{ \quad v[0], v[1], ., v[N-1], v[N], v[N+1], ., v[2N-1].. \quad \right\} \rightarrow \left\{ \begin{bmatrix} v[0] \\ v[1] \\ . \\ . \\ v[N-1] \end{bmatrix} \begin{bmatrix} v[N] \\ v[N+1] \\ . \\ . \\ v[2N-1] \end{bmatrix} \cdots \right\} \tag{A.4}$$

Discrete-time inverse lifting operator by a factor of N is defined by $\mathbb{L}_N^{-1}$ in the time domain and it is defined as [1]

$$\mathbb{L}_N^{-1}: \quad l^2(\mathbb{Z}, \mathbb{R}^N) \to l^2(\mathbb{Z}, \mathbb{R}), \tag{A.5}$$

$$\left\{ \begin{bmatrix} v_0[0] \\ v_1[0] \\ v_2[0] \\ . \\ . \\ . \\ v_{N-1}[0] \end{bmatrix} \begin{bmatrix} v_0[1] \\ v_1[1] \\ v_2[1] \\ . \\ . \\ . \\ v_{N-1}[1] \end{bmatrix} \ldots \right\} \to v_0[0], v_1[0] \ldots v_{N-1}[0], v_0[1], v_1[1] \ldots v_{N-1}[1] \ldots \tag{A.6}$$

The z-transform representations of lifting and inverse lifting are [47, 112],

$$\mathbf{L_N} = (\downarrow N) \begin{bmatrix} 1 & z & z^2 & \ldots.. & z^{N-1} \end{bmatrix}^T \tag{A.7a}$$

$$\mathbf{L_N^{-1}} = \begin{bmatrix} 1 & z^{-1} & z^{-2} & \ldots.. & z^{-(N-1)} \end{bmatrix} (\uparrow N). \tag{A.7b}$$

$\mathbf{L_N}$ and $\mathbf{L_N^{-1}}$ are denoting the z-transform of lifting and inverse lifting by a factor N, respectively. Lifting technique is time varying and non-causal in nature, and inverse lifting is causal and time varying in nature.

Proposition 1. *Let transfer function $\mathcal{F}(z)$ be represented in state space as*

$$\mathcal{F}(z) := \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] = D + C(zI - A)^{-1}B,$$

with $A \in \mathbb{R}^{N \times N}, B \in \mathbb{R}^{N \times p}, C \in \mathbb{R}^{m \times N}, D \in \mathbb{R}^{m \times p}$ matrices, m and p being the dimensions of output and input of $\mathcal{F}(z)$, respectively. Next, the lifted (by a factor of N) transfer function of

$\mathcal{F}(z)$ in state space form is represented as

$$\overline{\mathcal{F}}(z) := \mathbf{L_N}\mathcal{F}(z)\mathbf{L_N^{-1}} = \left[\begin{array}{c|cccccc} A^N & A^{N-1}B & A^{N-2}B & . & . & . & B \\ \hline C & D & 0 & 0 & 0 & 0 & 0 \\ CA & CB & D & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ CA^{N-1} & CA^{N-2}B & CA^{N-3}B & . & . & . & D \end{array}\right], \tag{A.8}$$

where $\mathbf{L_2}$ and $\mathbf{L_2^{-1}}$ can be obtained by using (A.7a) *and* (A.7b)*, respectively.*

Proof. See [1, Theorem 8.2.1]. □

B

A general solution of sampled-data system problem in ABE

A general sampled-data error system in ABE is shown in Figure B.1. The error system $\mathbb{G}$ can be written in z domain

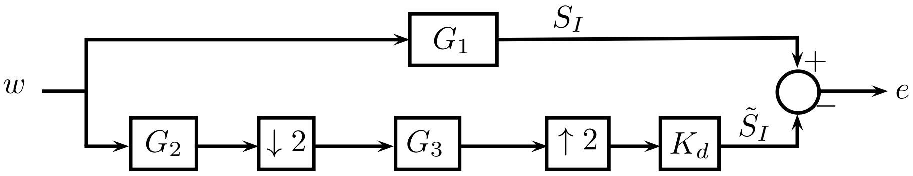

Figure B.1: A general sampled-data error system in ABE.

$$\mathbb{G}(z) = G_1(z) - K_d(z)(\uparrow 2)G_3(z)(\downarrow 2)G_2(z), \tag{B.1}$$

The error system $\mathbb{G}$ in Figure B.1 is a multi rate system because of the presence of the upsampler and downsampler. Hence, this system needs to be converted into a single rate system for obtaining the solution using the MATLAB robust control toolbox [113, 114]. System $\mathbb{G}$ can be transformed into a single rate system $\overline{\mathbb{G}}$ by using the lifting operation [1, 48], defined in (A.7). (A.8) is used to get the following results in [47]

$$\begin{aligned} K_d(z)(\uparrow 2) &= \mathbf{L_2^{-1}L_2}K_d(z)\mathbf{L_2^{-1}L_2}(\uparrow 2), \\ &= \mathbf{L_2^{-1}}\overline{K}_d(z)\begin{bmatrix}1 & 0\end{bmatrix}^T_{1\times 2}, \\ &= \mathbf{L_2^{-1}}\tilde{K}_d(z), \qquad (B.2) \\ K_d(z) &= \begin{bmatrix}1 & z^{-1}\end{bmatrix}\tilde{K}_d(z^2), \qquad (B.3) \end{aligned}$$

where

$$\tilde{K}_d(z) := \overline{K}_d(z)\begin{bmatrix}1 & 0\end{bmatrix}^T_{1\times 2}, \tag{B.4}$$

$$\overline{K}_d(z) := \mathbf{L_2}K_d(z)\mathbf{L_2^{-1}}. \tag{B.5}$$

Equality defined in (B.2) is substituted in (B.1) as

$$\mathbb{G}(z) = G_1(z) - \mathbf{L_2^{-1}}\tilde{K}_d(z)G_3(z)(\downarrow 2)G_2(z). \tag{B.6}$$

In (B.6), all the transfer functions do not have the same sampled rate, such as transfer functions $\tilde{K}_d(z)$ and $G_3(z)$ sampled at 8 kHz and transfer functions $G_1(z)$ and $G_2(z)$ sampled at 16 kHz, i.e., the system $\mathbb{G}$ is a multi rate system. It can be transformed into a single rate system using lifting and inverse lifting operations, as defined in (A.7) [1, 48]. For this, the lifting is applied to the input and output of system $\mathbb{G}$. This leads to a lifted transfer function of the system $\mathbb{G}$, which is defined as

$$\begin{aligned}
\overline{\mathbb{G}}(z) &= \mathbf{L_2}\mathbb{G}(z)\mathbf{L_2^{-1}}, \\
&= \mathbf{L_2}G_1(z)\mathbf{L_2^{-1}} - \mathbf{L_2}\mathbf{L_2^{-1}}\tilde{K}_d(z)G_3(z)(\downarrow 2)G_2(z)\mathbf{L_2^{-1}}, \\
&= \mathbf{L_2}G_1(z)\mathbf{L_2^{-1}} - \mathbf{L_2}\mathbf{L_2^{-1}}\tilde{K}_d(z)G_3(z)(\downarrow 2)\mathbf{L_2^{-1}}\mathbf{L_2}G_2(z)\mathbf{L_2^{-1}}, \\
&= \overline{G}_1(z) - \tilde{K}_d(z)G_3(z)S\overline{G}_2(z),
\end{aligned} \tag{B.7}$$

where $\mathbf{L_2}\mathbf{L_2^{-1}} = \mathbf{L_2^{-1}}\mathbf{L_2} = 1$, $\overline{G}_1(z) := \mathbf{L_2}G_1(z)\mathbf{L_2^{-1}}$, $S = \begin{bmatrix} 1 & 0 \end{bmatrix}$, and $\overline{G}_2(z) := \mathbf{L_2}G_2(z)\mathbf{L_2^{-1}}$. The lifted transfer function $\overline{\mathbb{G}}(z)$ is a single-rate system at 8 kHz. The H^∞-norm of the system $\overline{\mathbb{G}}(z)$ is equal to the H^∞-norm of the system $\mathbb{G}(z)$ as the lifting does not change the H^∞-norm [1]. The H^∞-norm of the system $\mathbb{G}$ is minimized using the optimal filter $\tilde{K}_d(z)$. Equation (B.7) can be written in the form of a standard feedback control system (closed-loop system) by using (A.1), as depicted in Figure B.2 [1]. Here, $\mathbf{0}$ is a zero matrix of 1×2, $\mathbf{I}$ is an identity matrix of 2×2, $\tilde{w} = \mathbf{L_2}w$, and $\tilde{e} = \mathbf{L_2}e$. Further, the optimal filter $\tilde{K}_d(z)$ is obtained with the help of robust control toolbox in MATLAB [114]. To this end, the optimal filter $K_d(z)$ is obtained from $\tilde{K}(z)$ by using (B.3).

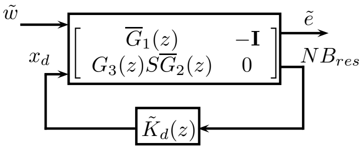

Figure B.2: General standard feedback control system.

C

Objective measures

Several standard objective speech quality measures such as mean square error (MSE) [75], signal to distortion ratio (SDR) [76], log likelihood ratio (LLR) [3, 77], logarithmic spectral distance LSD [39, 78], narrowband MOS-LQO (mean opinion score listening quality objective) [79, 80], and wideband MOS-LQO [81, 82] are computed for performance analysis. The mathematical formulation is written.

$$\text{MSE} = \frac{\sum_{i=1}^{L}(s(i) - \tilde{s}(i))^2}{L} \tag{C.1}$$

L is signal length, s is the original wideband signal, and $\tilde{s}$ is the reconstructed wideband signal.

$$\text{SDR(dB)} = 10\log_{10}\frac{\sum_{i=1}^{L}(s(i)^2}{\sum_{i=1}^{L}(s(i) - \tilde{s}(i))^2} \tag{C.2}$$

Parameters in (C.2) are the same as defined in (C.1).

$$\text{LLR} = \frac{\sum_{i=1}^{M}\log_{10}\big(\frac{\overrightarrow{a_i}_p^T R_{ic}\overrightarrow{a_{ip}}}{\overrightarrow{a_i}_c^T R_{ic}\overrightarrow{a_i}_\mathbf{c}}\big)}{M}. \tag{C.3}$$

M is the number of frames, $\overrightarrow{a_i}_c$ and $\overrightarrow{a_i}_p$ are the LPC vector of the original i^{th} speech frame and reconstructed i^{th} speech frame, respectively, and R_{ic} is the autocorrelation matrix of the original i^{th} speech frame.

$$\text{LSD} = \frac{\sum_{i=1}^{M}\sqrt{\big(\frac{\sum_{j=n_{low}}^{n_{high}}(20\log_{10}|X(i,j)|-20\log_{10}|\tilde{X}(i,j)|)^2}{N}\big)}}{M} \tag{C.4}$$

with $|X(i,j)|$ and $\tilde{X}(i,j)$ being the absolute values of the FFT of i^{th} frame and j^{th} frequency bin of original and reconstructed speech frame, respectively. n_{low} and n_{high} are the frequency bins corresponding to the frequency range from 0 or 4 to 7 or 8 kHz. M and N are denoting the number of frames and the number of frequency bins, respectively.

$$\text{MOS-LQO} = a + \frac{b}{(1 + \exp(c * p + d))} \tag{C.5}$$

with $a = 0.999$, $b = 4.999 - a$, $c = -1.4945$ for narrowband MOS-LQO and $= -1.3669$ for wideband MOS-LQO, $d = 4.6607$ for narrowband MOS-LQO and $= 3.8224$ for wideband MOS-LQO, and p is PESQ. PESQ measure is used reliably to predict the speech quality in a wider range of network conditions, including analog connections, codecs, packet loss, and variable delay. PESQ measuring process consists of the level alignment of the original signal and reconstructed signal to a standard listening level, filtering process, time alignment for correcting time delays, auditory transform process to obtain the loudness spectra, calculating the difference between the loudness spectra, and averaging over time and frequency [3].

LLR, SDR, and narrowband PESQ measures are computed with the help of a composite tool downloaded from the website of the author, and the narrowband MOS-LQO measure is computed from the narrowband PESQ [79,80]. The wideband MOS-LQO measure is computed by the MATLAB function *PESQ2_MTLB* downloaded from the mathworks website [82].

Bibliography

[1] T. Chen and B. A. Francis, *Optimal sampled-data control systems.* Springer, 1995, vol. 124.

[2] "ITU-T Software Tool Library 2009 Users Manual," *ITU-T Recommendation G.191*, Nov. 2009.

[3] P. C. Loizou, *Speech enhancement: theory and practice*, 2nd ed. CRC press, 2007.

[4] X. Shao, "Robust Algorithms for Speech Reconstruction on Mobile Devices," Ph.D. dissertation, University of East Anglia, 2005.

[5] J. Makhoul, "Linear prediction: A tutorial review," *in Proceedings IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[6] L. Laaksonen *et al.*, "Artificial bandwidth extension of narrowband speech-enhanced speech quality and intelligibility in mobile devices," 2013.

[7] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Cambridge, United Kingdom*, vol. 4, 1979, pp. 428–431.

[8] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the Mel frequency cepstral coefficients," in *Proceedings IEEE Workshop on Speech Coding*, 1999, pp. 171–173.

[9] N. Prasad and T. K. Kumar, "Bandwidth Extension of Speech Signals: A Comprehensive Review," *International Journal of Intelligent Systems and Applications*, vol. 8, no. 2, p. 45, 2016.

[10] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 1, pp. 266–274, 2001.

[11] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[12] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 86, no. 6, pp. 1296–1306, 2006.

[13] J. Abel and T. Fingscheidt, "Artificial Speech Bandwidth Extension Using Deep Neural Networks for Wideband Spectral Envelope Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2018.

[14] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech," in *Eighth European Conference on Speech Communication and Technology, GENEVA, Switzerland*, 2003, pp. 1433–1436.

[15] I. Soon and C. Yeo, "Bandwidth extension of narrowband speech using cepstral analysis," in *Proceedings of IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004, pp. 242–245.

[16] Y. Li and S. Kang, "Artificial bandwidth extension using deep neural network-based spectral envelope estimation and enhanced excitation estimation," *IET Signal Processing*, vol. 10, no. 4, pp. 422–427, 2016.

[17] B. Andersen, J. Dyreby, B. Jensen, F. H. Kjærskov, O. L. Mikkelsen, P. D. Nielsen, and H. Zimmermann, "Bandwidth Expansion of Narrow Band Speech using Linear Prediction," *web source*, vol. 26, 2015.

[18] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 1, 2001, pp. 665–668.

[19] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2005, pp. I–805.

[20] H. Pulakka, U. Remes, K. Palomäki, M. Kurimo, and P. Alku, "Speech bandwidth extension using Gaussian mixture model-based estimation of the highband mel spectrum," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5100–5103.

[21] A. H. Nour-Eldin and P. Kabal, "Memory-based approximation of the Gaussian mixture model framework for bandwidth extension of narrowband speech," in *Proceedings Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[22] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, "GMM-based bandwidth extension using sub-band basis spectrum model," in *Proceedings Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[23] P. B. Bachhav, M. Todisco, M. Mossi, C. Beaugeant, and N. Evans, "Artificial bandwidth extension using the constant-Q transform," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5550–5554.

[24] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5029–5033.

[25] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden Markov model," in *Proceedings IEEE Workshop on Speech Coding*, 2000, pp. 133–135.

[26] M. L. Seltzer and A. Acero, "Training wideband acoustic models using mixed-bandwidth training data for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 235–245, 2007.

[27] G.-B. Song and P. Martynovich, "A study of HMM-based bandwidth extension of speech signals," *Signal Processing*, vol. 89, no. 10, pp. 2036–2044, 2009.

[28] C. YağLı, M. T. Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a viterbi path," *Speech Communication*, vol. 55, no. 1, pp. 111–118, 2013.

[29] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal processing magazine*, vol. 29, 2012.

[30] 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, 2014.

[31] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, “Speech bandwidth expansion based on deep neural networks,” in *Proceedings Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[32] J. Abel and T. Fingscheidt, “A dnn regression approach to speech enhancement by artificial bandwidth extension,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 219–223.

[33] K.-T. Kim, M.-K. Lee, and H.-G. Kang, “Speech bandwidth extension using temporal envelope modeling,” *IEEE Signal Processing Letters*, vol. 15, pp. 429–432, 2008.

[34] Y. Sunil and R. Sinha, “Sparse representation based approach to artificial bandwidth extension of speech,” in *2014 International Conference on Signal Processing and Communications (SPCOM)*, 2014, pp. 1–5.

[35] H. Tolba and D. O’Shaughnessy, “On the application of the AM-FM model for the recovery of missing frequency bands of telephone speech,” in *Fifth International Conference on Spoken Language Processing, Sydney, Australia*, 1998.

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

[37] L. Bin, T. Jianhua, W. Zhengqi, L. Ya, D. Bukhari *et al.*, “A novel method of artificial bandwidth extension using deep architecture,” 2015.

[38] J. Sadasivan, S. Mukherjee, and C. S. Seelamantula, “Joint dictionary training for bandwidth extension of speech signals,” in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5925–5929.

[39] J. Abel, M. Strake, and T. Fingscheidt, “A simple cepstral domain DNN approach to artificial speech bandwidth extension,” in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5469–5473.

[40] D. Marelli and P. Balazs, “On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 237–248, 2010.

[41] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, “State-space solutions to standard H^2 and H^∞ control problems,” *IEEE Transactions on Automatic control*, vol. 34, no. 8, pp. 831–847, 1989.

[42] Y. Yamamoto, “A function space approach to sampled data control systems and tracking problems,” *IEEE Transactions on Automatic Control*, vol. 39, no. 4, pp. 703–713, 1994.

[43] S. Ashida, M. Nagahara, and Y. Yamamoto, “Audio signal compression via sampled-data control theory,” in *SICE 2003 Annual Conference (IEEE Cat. No. 03TH8734)*, vol. 2, 2003, pp. 1744–1747.

[44] Z. Du, Z. Yan, and Z. Zhao, “Interval type-2 fuzzy tracking control for nonlinear systems via sampled-data controller,” *Fuzzy Sets and Systems*, vol. 356, pp. 92–112, 2019.

[45] Z. Du, Y. Kao, and J. H. Park, “New results for sampled-data control of interval type-2 fuzzy nonlinear systems,” *Journal of the Franklin Institute*, vol. 357, no. 1, pp. 121–141, 2020.

[46] H. S. Shekhawat and G. Meinsma, “A sampled-data approach to optimal non-causal downsampling,” *Mathematics of Control, Signals, and Systems*, vol. 27, no. 3, pp. 277–315, 2015.

[47] Y. Yamamoto, M. Nagahara, and P. P. Khargonekar, “Signal Reconstruction via H^{∞} Sampled-Data Control Theory Beyond the Shannon Paradigm,” *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 613–625, 2012.

[48] T. Chen and B. A. Francis, “Design of multirate filter banks by H^{∞} optimization,” *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2822–2830, 1995.

[49] Y. Yamamoto, H. Fujioka, and P. P. Khargonekar, “Signal reconstruction via sampled-data control with multirate filter banks,” in *Proceedings 36th IEEE Conference on Decision and Control*, vol. 4, 1997, pp. 3395–3400.

[50] Y. Yamamoto, M. Nagahara, and H. Fujioka, “Multirate Signal Reconstruction and Filter Design Via Sampled-Data Control,” *MTNS*, 2000.

[51] Z. Du, Y. Kao, H. R. Karimi, and X. Zhao, “Interval Type-2 Fuzzy Sampled-Data H^{∞} Control for Nonlinear Unreliable Networked Control Systems,” *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 7, pp. 1434–1448, 2019.

[52] U. Shaked and Y. Theodor, “H^{∞} optimal estimation: a tutorial,” in *Proceedings 31st IEEE Conference on Decision and Control*, 1992, pp. 2278–2286.

[53] J. Abel, M. Strake, and T. Fingscheidt, “Artificial bandwidth extension using deep neural networks for spectral envelope estimation,” in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[54] W. Nogueira, J. Abel, and T. Fingscheidt, “Artificial speech bandwidth extension improves telephone speech intelligibility and quality in cochlear implant users,” *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1640–1649, 2019.

[55] A. H. Nour-Eldin and P. Kabal, “Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech,” in *Proceedings Ninth Annual Conference of the International Speech Communication Association*, 2008.

[56] “EVS Permanent Document EVS-7c: Processing Functions for Characterization Phase (3GPP S4 141126, V. 1.0.0),” Aug. 2014.

[57] D. Gupta and H. S. Shekhawat, “Artificial bandwidth extension using H^{∞} optimization,” *Proc. Interspeech 2019*, pp. 3421–3425, 2019.

[58] D. Gupta, H. S. Shekhawat, and R. Sinha, “A new framework for artificial bandwidth extension using H^{∞} filtering,” *Circuits, Systems, and Signal Processing*, pp. 1–25, 2022, https://rdcu.be/cFTQQ.

[59] G. Meinsma and L. Mirkin, “Sampling from a system-theoretic viewpoint: Part Iconcepts and tools,” *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3578–3590, 2010.

[60] D. Gupta and H. Shekhawat, "Artificial Bandwidth Extension Using H^∞ Optimization and Speech Production Model," in *29th IEEE International Conference Radioelektronika (RADIOELEKTRONIKA)*, 2019, pp. 1–6.

[61] D. Gupta and H. S. Shekhawat, "High-band feature extraction for artificial bandwidth extension using deep neural network and H^∞ optimisation," *IET Signal Processing*, vol. 14, no. 10, pp. 783–790, 2021.

[62] J. D. Markel and A. G. Jr., *Linear Prediction of Speech*, 1st ed., ser. Communication and Cybernetics 12. Springer-Verlag Berlin Heidelberg, 1976.

[63] MathWorks, "http://www.mathworks.com/."

[64] K. Aida-Zade, C. Ardil, and S. Rustamov, "Investigation of combined use of MFCC and LPC features in speech recognition systems," *World Academy of Science, Engineering and Technology*, vol. 19, pp. 74–80, 2006.

[65] F. Itakula, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signal," *Journal of Acoustic Society of America*, 1975.

[66] Y. Sunil and R. Sinha, "Exploration of class specific ABWE for robust children's ASR under mismatched condition," in *Proceedings International Conference on Signal Processing and Communications (SPCOM)*, 2012, pp. 1–5.

[67] M. B. Christopher, *Pattern recognition and machine learning.* Springer-Verlag New York, 2016.

[68] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1998, pp. 285–288.

[69] H. V. Poor, *An introduction to signal detection and estimation.* Springer Science & Business Media, 2013.

[70] A. C. Ian Goodfellow, Yoshua Bengio, *Deep Learning.* MIT Press, 2016.

[71] W. Verhelst, "Overlap-add methods for time-scaling of speech," *Speech Communication*, vol. 30, no. 4, pp. 207–221, 2000.

[72] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.

[73] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[74] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[75] P. Nizampatnam and K. K. Tappeta, "Bandwidth extension of narrowband speech using integer wavelet transform," *IET Signal Processing*, vol. 11, no. 4, pp. 437–445, 2016.

[76] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Detection, separation and recognition of speech from continuous signals using spectral factorisation," in *Proceedings 20th IEEE European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2649–2653.

[77] 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 *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[78] J. Abel, M. Kaniewska, C. Guillaumé, W. Tirry, and T. Fingscheidt, "An instrumental quality measure for artificially bandwidth-extended speech signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 384–396, 2016.

[79] "ITU-T, Recommendation P862.1: Mapping function for transforming P. 862 raw result scores to MOS-LQO," *International Telecommunication Union, Geneva, Switzerland*, 2003.

[80] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[81] " ITU-T (2005), P.862 Amendment 2: Revised Annex A - Reference implementations and conformance testing for ITU-T Recs P.862, P.862.1 and P.862.2, http://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/en," *ITU-T Recommendation.*

[82] K. Wojcicki, "PESQ MATLAB Wrapper, https://www.mathworks.com/matlabcentral/fileexchange/33820-pesq-matlab-wrapper," *MATLAB Central File Exchange*, June 12, 2020.

[83] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[84] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469–472, 2009.

[85] N. Adiga and S. Prasanna, "Detection of glottal activity using different attributes of source information," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2107–2111, 2015.

[86] "ITU-T, Recommendation P.800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, p. 22, 1996.

[87] D. Gupta and H. S. Shekhawat, "Artificial bandwidth extension using deep neural network and H^∞ sampled-data control theory," *arXiv preprint arXiv:2108.13326*, 2021.

[88] ——, "Artificial bandwidth extension using H^∞ sampled-data control theory," *Speech Communication*, vol. 134, pp. 32–41, 2021, https://doi.org/10.1016/j.specom.2021.08.004.

[89] "Mandatory speech codec speech processing functions: Adaptive Multi-rate (AMR) speech codec; transcoding fucntions, 3GPP TS 26.090 Rel. 8," 2008.

[90] "ITU-T, Recommendation P. 56, Objective Measurement of Active Speech Level," *International Telecommunication Union*, 2011.

[91] J. D. Markel and A. J. Gray, *Linear prediction of speech.* Springer Science & Business Media, 2013, vol. 12.

[92] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband Mel spectrum," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, 2011.

[93] B. Iser, G. Schmidt, and W. Minker, *Bandwidth extension of speech signals.* Springer Science & Business Media, 2008, vol. 13.

[94] J. Zhang, L. Chai, C. Zhang, and E. Mosca, "Multi-objective approximation of IIR by FIR digital filters," in *6th IEEE World Congress on Intelligent Control and Automation*, vol. 2, 2006, pp. 6574–6577.

[95] K. Steiglitz, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 229–234, 1977.

[96] K. Schnell and A. Lacroix, "Pole zero estimation from speech signals by an iterative procedure," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 1, 2001, pp. 109–112.

[97] T. Kobayashi and S. Imai, "Design of iir digital filters with arbitrary log magnitude function by wls techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 247–252, 1990.

[98] M. A. Blommer and G. H. Wakefield, "On the design of pole-zero approximations using a logarithmic error measure," *IEEE Transactions on Signal processing*, vol. 42, no. 11, pp. 3245–3248, 1994.

[99] C. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[100] W. Sun, K. Nagpal, and P. Khargonekar, "H_∞ control and filtering for sampled-data systems," *IEEE Transaction of Automatic Control*, vol. 38, no. 8, pp. 1162–1175, August 1993.

[101] M. Unser, "optimality of ideal filters for pyramid and wavelet signal approximation," *IEEE Transaction on Signal Processing*, vol. 41, no. 12, pp. 3591 –3596, dec. 1993.

[102] P. P. Khargonekar and Y. Yamamoto, "Delayed signal reconstruction using sampled-data control," in *Proceedings 35th IEEE Conference on Decision and Control*, vol. 2, 1996, pp. 1259–1263.

[103] H. Ishii, Y. Yamamoto, and B. A. Francis, "Sample-rate conversion via sampled-data H^∞ control," in *Proceedings 38th IEEE Conference on Decision and Control*, vol. 4, 1999, pp. 3440–3445.

[104] M. Nagahara, "Multirate Digital Signal Processing via Sampled-Data H^∞ Optimization," Ph.D. dissertation, Kyoto University, 2003.

[105] M. Nagahara and Y. Yamamoto, "A new design for sample-rate converters," in *IEEE Conference on Decision and Control*, vol. 5, 2000, pp. 4296–4301.

[106] H. Kakemizu, M. Nagahara, A. Kobayashi, and Y. Yamamoto, "Noise reduction of jpeg images by sampled-data H^∞ optimal ε filters," in *Proceedings SICE Annual Conference*, 2005, pp. 1080–1085.

[107] Y. Yamamoto, M. Nagahara, and P. Khargonekar, "A Brief Overview of Signal Reconstruction via Sampled-Data H^∞ Optimization," *Applied and Computational Mathematics*, vol. 11, no. 1, pp. 3 –18, 2012.

[108] G. Meinsma and L. Mirkin, "Sampling from a System-Theoretic Viewpoint: Part II—Non-causal Solutions," *IEEE Transaction on Signal Processing*, vol. 58, no. 7, pp. 3591–3606, July 2010.

[109] ——, "L^2 Sampled signal reconstruction with causality constraints - Part I: Setup and solutions," *IEEE Transaction on Signal Processing*, vol. 60, no. 5, pp. 2260–2272, 2012.

[110] ——, "L^2 Sampled Signal Reconstruction With Causality Constraints - Part II: Theory," *IEEE Transaction on Signal Processing*, vol. 60, no. 5, pp. 2273–2285, 2012.

[111] H. S. Shekhawat and G. Meinsma, "A sampled-data approach to optimal relaxed-causal sampling," *Mathematics of Control, Signals and Systems*, 2021, https://doi.org/10.1007/s00498-021-00297-9.

[112] P. P. Vaidyanathan, *Multirate systems and filter banks*, ser. Prentice-Hall signal processing series. Prentice Hall, 1993.

[113] K. Glover and J. C. Doyle, "State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relations to relations to risk sensitivity," *Systems & control letters*, vol. 11, no. 3, pp. 167–172, 1988.

[114] R. Y. Chiang and M. G. Safonov, *MATLAB : Robust Control Toolbox User's Guide.* Math Works, 1997.

Publications during thesis work

- Journals

  (i) Deepika Gupta, H. S. Shekhawat, "High-band Feature Extraction for Artificial Bandwidth Extension Using Deep Neural Network and H^{∞} optimization", *IET Signal Processing* IET Signal Processing, volume 14, no. 10, pp. 783790, 2020, https://doi.org/10.1049/iet-spr.2020.0214.

  (ii) Deepika Gupta, H. S. Shekhawat, "Artificial Bandwidth Extension Using H^{∞} Sampled-data control theory", *Speech Communication, Elsevier*, volume 134, pp. 32-41, 2021, https://doi.org/10.1016/j.specom.2021.08.004.

  (iii) Deepika Gupta, H. S. Shekhawat and Rohit Sinha, "A New Framework for Artificial Bandwidth Extension using H^{∞} Filtering", *Circuits, Systems, and Signal Processing, Springer*, pp. 1-25, 2022, https://rdcu.be/cFTQQ.

  (iv) Deepika Gupta, H. S. Shekhawat, "Artificial Bandwidth Extension using Frequency Shifting, H^{∞} Optimization, and Deep Neural Network", *communicated*.

- Conferences

  (i) Deepika Gupta and H. S. Shekhawat,"Artificial Bandwidth Extension Using H^{∞} Optimization and Speech Production Model", in Proceedings 29th International Conference Radioelektronika (RADIOELEKTRONIKA), MAREW 2019, https://doi.org/10.1109/RADIOELEK.2019.8733452.

  (ii) Deepika Gupta and H. S. Shekhawat, "Artificial Bandwidth Extension using H^{∞} Optimization", in Proceedings Interspeech 2019, http://dx.doi.org/10.21437/Interspeech.2019-1580.