ARTIFICIAL BANDWIDTH EXTENSION OF SPEECH USING TEMPORAL CLUSTERING

by

Can Yağlı

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Electrical & Computer Engineering

Koç University

November, 2010

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Can Yağlı

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assoc. Prof. Engin Erzin

Assoc. Prof. Yücel Yemez

Prof. Levent M. Arslan

Date:

To my beloved family

ABSTRACT

Historically public telephone networks operate with narrowband speech, which is bandlimited to (250, 3400) Hz in frequency. Even though public telephone exchanges are digital today, the low bandwidth limitation is still present due to the characteristics of the traditional analogue network and related standards. Although intelligibility of the narrowband speech is high, studies show that the perceived quality of the narrowband speech is significantly degraded compared to wideband speech, which is bandlimited to (50, 7000) Hz in frequency. In this thesis, we investigate the Artificial Bandwidth Extension problem, which aims to reconstruct the missing frequency in wideband speech from narrowband speech. To solve the problem, we utilize the well-known source-filter reproduction of the human voice production system. This model decomposes the speech signal into two, namely the source signal and the filter representing spectral envelope. The source signal is extended with up-sampling with zero insertion (spectral folding) and we propose a new framework for the estimation of wideband spectral envelope from narrowband. The proposed framework builds temporal clusters of the joint sub-phone patterns of the narrowband and wideband speech signals using a parallel branch HMM structure. The joint sub-phone patterns define temporally correlated neighborhoods, in which a linear prediction filter estimates spectral features of the corresponding wideband signal from the narrowband signal. The proposed framework is compared to a benchmark vector quantization based artificial bandwidth extension algorithm. Objective metrics and a subjective test shows that the reconstructed wideband speech with our method significantly outperforms the narrowband speech and wideband speech reconstructed with the benchmark system.

ÖZETÇE

Analog şehir hatlarında iletilen telefon sinyali frekans bölgesinde 250-3400 Hz ile sınırlı iken (dar bant konuşma) modern dijital hatlarda bu frekans üst sınırı 8000 Hz olarak belirlenmiştir (geniş bant konuşma). Her ne kadar dar bant konuşmanın anlaşılabilirliği yeterince yüksek olsa da insan kulağına geniş bant konuşma kadar hoş gelmediği tespit edilmiştir. Yapay Bant Genişliği Artırma, dar bant konuşmadan geniş bant konuşmaya geçişi amaçlar. Bu tezin amacı da bu probleme bir çözüm getirmektir. Problemin çözümü için Kaynak-Süzgeç modelinden faydalanılmıştır. Bu model, ses sinyalini kaynak ve süzgeç olarak ikiye ayırır. Kaynak sinyali spektral katlama ile genişletilirken süzgeç genişletilmesi için yeni bir yöntem önerilmektedir. Önerilen yöntem paralel yapılı Gizli Markov Modellerinden faydalanarak dar bant ve geniş bant konuşma sinyalleri arasındaki ortak benzerlikleri zamansal olarak gruplandırır. Bu gruplar kullanılarak dar banttan geniş banta geçişi sağlayan doğrusal süzgeçler elde edilir. Önerilen yöntem vektör nicemleme kullanan temel bir yöntemle karşılaştırılmıştır. Nesnel ve öznel sınamalar önerilen yöntemin dar bant konuşmayı da vektör nicemleme kullanan yöntemi de alt ettiğini göstermektedir.

ACKNOWLEDGMENTS

Firstly, I would like thank my advisor Prof. Engin Erzin who has guided me through my M.Sc. studies and provided invaluable support.

I would also like to thank Prof. Yücel Yemez and Prof. Levent Arslan who are the other members of my thesis committee.

Finally, I am grateful to my family and my friends who supported and motivated me even during the hard times.

TABLE OF CONTENTS

List of Tables ix

List of Figures xi

Nomenclature xiii

Chapter 1: Introduction 1

1.1 Previous Work 2

1.2 Contributions 5

1.3 Outline 5

1.4 Notation 6

Chapter 2: Speech Production System and Benchmark System 7

2.1 Source-Filter Separation 7

2.1.1 Source Signal 8

2.1.2 Filter 8

2.2 Linear Predictive Coding 9

2.2.1 Line Spectrum Frequencies 10

2.3 Benchmark VQ System 11

2.3.1 Extension of the excitation 11

2.3.2 Extension of the envelope 11

2.3.3 Summary of VQ Algorithm 13

Chapter 3: Proposed Model for Artificial Bandwidth Extension of the Envelope 14

3.1 Markov Models 14

3.2 Hidden Markov Models 15

3.2.1 Hidden Markov Model Parameters 17
3.3 Proposed Model 18
3.3.1 Temporal Clustering 18
3.3.2 Linear Estimation of the Wideband Spectra 19
3.3.3 Feature Selection 20
3.3.4 Reducing Number of Filters 21
3.4 Summary of the Proposed Method 21

Chapter 4: Experimental Results 23

4.1 The TIMIT Database 23
4.2 Objective Metrics 24
4.2.1 Segmental Signal-to-Noise Ratio 24
4.2.2 Logarithmic Spectral Distortion 24
4.2.3 Perceptual Evaluation of Speech Quality 25
4.3 Performance of the Benchmark VQ System 25
4.4 Performance of the Proposed System 26
4.5 Discussion of the Proposed System 62
4.5.1 Comparison with the Benchmark System 62
4.5.2 The Effect of the Branch Number B 63
4.5.3 The Effect of Temporal Neighborhood 63
4.5.4 The Effect of Feature Selection 64
4.5.5 The Effect of Index Determinition 65
4.6 Subjective Test Results and Evaluation 65

Chapter 5: Conclusion 67

5.1 Future Work 69

Bibliography 70

LIST OF TABLES

4.1 The Structure of Our TIMIT Database. 23

4.2 Performance of the VQ based benchmark ABE system. 25

4.3 SegSNR scores for SA set, $T = 0$, indexes are saved for every state s_b 27

4.4 SegSNR scores for SX set, $T = 0$, indexes are saved for every state s_b 28

4.5 LSD scores for SA set, $T = 0$, indexes are saved for every state s_b 29

4.6 LSD scores for SX set, $T = 0$, indexes are saved for every state s_b 31

4.7 PESQ scores for SA set, $T = 0$, indexes are saved for every state s_b 32

4.8 PESQ scores for SX set, $T = 0$, indexes are saved for every state s_b 34

4.9 SegSNR scores for SA set, $T = 1$, indexes are saved for every state s_b 35

4.10 SegSNR scores for SX set, $T = 1$, indexes are saved for every state s_b 37

4.11 LSD scores for SA set, $T = 1$, indexes are saved for every state s_b 38

4.12 LSD scores for SX set, $T = 1$, indexes are saved for every state s_b 40

4.13 PESQ scores for SA set, $T = 1$, indexes are saved for every state s_b 41

4.14 PESQ scores for SX set, $T = 1$, indexes are saved for every state s_b 43

4.15 SegSNR scores for SA set, $T = 0$, single index saved 44

4.16 SegSNR scores for SX set, $T = 0$, single index saved 46

4.17 LSD scores for SA set, $T = 0$, single index saved 47

4.18 LSD scores for SX set, $T = 0$, single index saved 49

4.19 PESQ scores for SA set, $T = 0$, single index saved 50

4.20 PESQ scores for SX set, $T = 0$, single index saved 52

4.21 SegSNR scores for SA set, $T = 1$, single index saved 53

4.22 SegSNR scores for SX set, $T = 1$, single index saved 55

4.23 LSD scores for SA set, $T = 1$, single index saved 56

4.24 LSD scores for SX set, $T = 1$, single index saved 58

4.25 PESQ scores for SA set, $T = 1$, single index saved 59

4.26 PESQ scores for SX set, $T = 1$, single index saved 61
4.27 The effect of temporal neighborhood T 64
4.28 The subjective A/B pair comparison test results. 66

LIST OF FIGURES

1.1 Wideband and Narrowband versions of the same speech signal 1

1.2 Artificial Bandwidth Extension with Source-Filter model 3

2.1 Voice Production System 7

2.2 Excitation and Envelope Signals 9

3.1 A Markov chain with 3 states and corresponding state transition probabilities 15

3.2 A Hidden Markov process with 3 states s_i, state transition probabilities a_{ij} and observable outputs Θ_i 16

3.3 Parallel HMM structure, where s_s and s_e are non-emitting states. 19

3.4 Proposed envelope extension method. 22

4.1 (a)Experimental Results for the SA set (b)Experimental Results for the SX set 26

4.2 SegSNR scores for SA set, $T = 0$, indexes are saved for every state s_b 27

4.3 SegSNR scores for SX set, $T = 0$, indexes are saved for every state s_b 29

4.4 LSD scores for SA set, $T = 0$, indexes are saved for every state s_b 30

4.5 LSD scores for SX set, $T = 0$, indexes are saved for every state s_b 32

4.6 PESQ scores for SA set, $T = 0$, indexes are saved for every state s_b 33

4.7 PESQ scores for SX set, $T = 0$, indexes are saved for every state s_b 35

4.8 SegSNR scores for SA set, $T = 1$, indexes are saved for every state s_b 36

4.9 SegSNR scores for SX set, $T = 1$, indexes are saved for every state s_b 38

4.10 LSD scores for SA set, $T = 1$, indexes are saved for every state s_b 39

4.11 LSD scores for SX set, $T = 1$, indexes are saved for every state s_b 41

4.12 PESQ scores for SA set, $T = 1$, indexes are saved for every state s_b 42

4.13 PESQ scores for SX set, $T = 1$, indexes are saved for every state s_b 44

4.14 SegSNR scores for SA set, $T = 0$, single index saved 45

4.15 SegSNR scores for SX set, $T = 0$, single index saved 47

4.16 LSD scores for SA set, $T = 0$, single index saved 48

4.17 LSD scores for SX set, $T = 0$, single index saved 50

4.18 PESQ scores for SA set, $T = 0$, single index saved 51

4.19 PESQ scores for SX set, $T = 0$, single index saved 53

4.20 SegSNR scores for SA set, $T = 1$, single index saved 54

4.21 SegSNR scores for SX set, $T = 1$, single index saved 56

4.22 LSD scores for SA set, $T = 1$, single index saved 57

4.23 LSD scores for SX set, $T = 1$, single index saved 59

4.24 PESQ scores for SA set, $T = 1$, single index saved 60

4.25 PESQ scores for SX set, $T = 1$, single index saved 62

4.26 Sample R_{yx} correlation matrix visualizations. Columns represent the narrowband LSF features and rows represent wideband LSF features. The intensity of the gray level represents higher correlation. 65

NOMENCLATURE

ABE	Artificial Bandwidth Extension
eb	Extension Band
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LPC	Linear Predictive Coding
LSD	Logarithmic Spectral Distortion
LSF	Line Spectral Frequencies
LSP	Line Spectral Pairs
SegSNR	Segmental Signal-to-Noise Ratio
nb	Narrowband
wb	Wideband
VQ	Vector Quantization

Chapter 1

INTRODUCTION

Conventionally, public telephone networks have allowed the operation of narrowband speech only, which is bandlimited between 250-3400 Hz in frequency, whereas digital networks support wideband speech that is bandlimited to 8 kHz. Despite the fact that the majority of telephone exchanges are conducted digitally today, this limitation on the bandwidth still applies thanks to the characteristics of the old analogue telephone networks and related standards. The bandlimited (i.e. narrowband) speech provides a nearly perfect intelligibility ratio, as [1] states; but other studies prove that the missing upper frequency band improves the perception of the speech quality significantly [2]. Therefore, presence of spectrally rich information in the upper frequency bands reduces the listening effort as well.

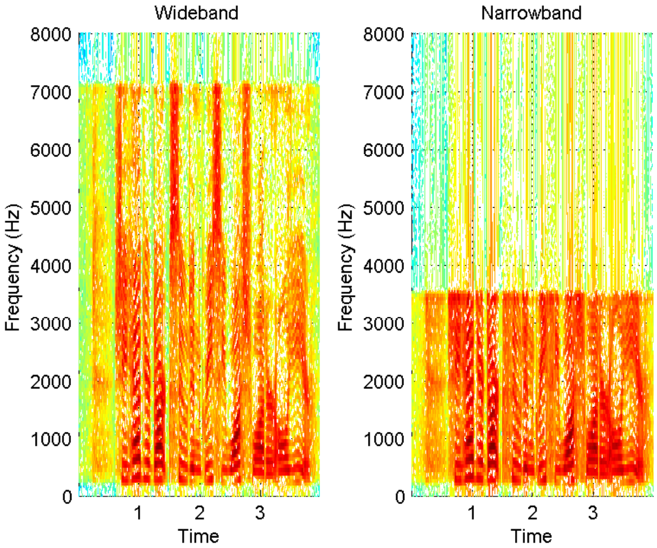


Figure 1.1: Wideband and Narrowband versions of the same speech signal

In Fig. 1.1 we see the wideband and narrowband versions of the same speech signal.

An important amount of important is missing in the narrowband case compared to the wideband.

The upgrading of the traditional analogue networks to support wideband speech is not likely in the near future, since such an infrastructural operation will be too costly. Instead, modifications on the received narrowband speech on the telephone end tend to be simpler and faster solutions to the problem. Artificial bandwidth extension, for this purpose, works on the estimation of the missing 3400 - 8000 Hz band from the narrowband speech.

In the literature and during this thesis, narrowband (or bandlimited) speech refers to the speech with spectral information only in 250-3400 Hz band. For simplicity, this interval is considered to be 0-4 kHz; or narrowband speech is known to be sampled with 8 kHz, as well. Wideband speech, on the other hand, contains spectral components from 0 up to 8 kHz, therefore is defined to be sampled with 16kHz. Artificial bandwidth extension can, then, be defined as mapping from narrowband speech to wideband speech. In addition, we will call the 4-8 kHz range as the extension band.

Artificial bandwidth extension improves the quality of the perceived sound. In addition, it can be implemented on the already existing analogue networks at the telephone end, therefore is helpful to avoid the economic burden of re-designing and constructing the network infrastructure.

1.1 Previous Work

Formerly, artificial bandwidth extension methods were elementary methods which exploited simple signal processing algorithms. These methods were generally called 'non-model based algorithms', as they did not benefit from the Source-Filter separation model (explained in Chapter 2). To name an example, we can refer to [3] which was used by BBC to improve the speech whilst broadcasting.

Most of the ABE methods today use the Source-Filter separation when analyzing speech signals. Simply, Source-Filter model decomposes the speech signal into two, namely the *excitation* (source) and the *envelope* (filter). In this framework, the ABE problem is divided into two as well, as the extension of the excitation and extension of the envelope, which is shown in Fig 1.2. Previous work [4] shows the extension of the excitation plays a minor role compared to the extension of the envelope in terms of improving the perceived speech quality and this is also apparent in the literature as well, since most of the endeavor is placed upon the methods that extend the envelope.

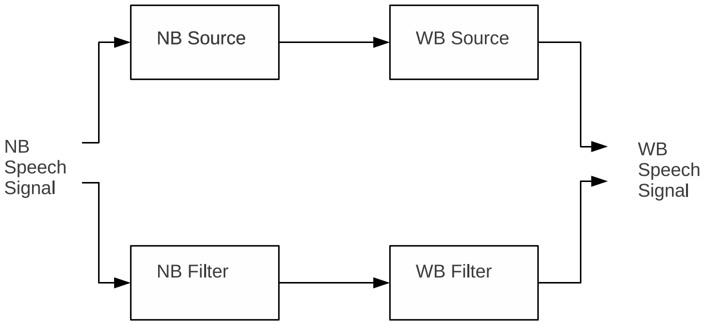


Figure 1.2: Artificial Bandwidth Extension with Source-Filter model

ABE algorithms try to define a correlation between the narrowband features and the wideband features (or, the extension band features). For this purpose, Enbom and Kleijn have used vector quantization [5]. Initially, for the training phase, they obtain a database that contains both narrowband and wideband versions of the speech signals. Utilizing the Source-Filter model, they access the excitation and envelope for each speech signal pair. They, then, extract the features that define these narrowband and wideband signals from the envelope. As the envelope-defining features, they use Mel Frequency Cepstral Coefficients (MFCCs). Fusing the narrowband and wideband features for a given speech frame together, they form a new vector for every frame. Finally, with these vectors, they construct codebooks with Linde-Buzo-Gray algorithm [6]. During the application phase, the features of the narrowband test signal is compared to the narrowband parts of the vectors in the codebook. The nearest vector in terms of euclidian distance is chosen. The wideband part of this very vector is decided as the wideband correspondent of the narrowband test signal.

Unlike Enbom and Kleijn, Kim and Park use Gaussian Mixture Methods instead of

VQ in order to estimate the wideband envelope from the narrowband envelope [7]. Similar to [5], the method is initialized by constructing the fused vector database, which contains the packed and concatenated narrowband and wideband features. Following the initialization, Kim and Park use GMMs to model, group and parameterize these vectors. The training phase is completed when the parameters for all models are obtained. Therefore, whenever a narrowband test signal is introduced to the system,

1. Its coefficients with respect to the available Gaussian models are calculated.
2. Using these coefficients and the models, the wideband features are found with a straight forward manner.

Even though VQ and GMM are efficient machine learning algorithms, they do not focus on the correlation among subsequent frames. In other words, while calculating the wideband correspondent of a given narrowband frame, no information from the previous frame is used, so these algorithms do not benefit from temporal information. Jax and Vary [4, 8] suggest usage of Hidden Markov Models to define the correlation between narrowband and wideband features, and this incorporates the temporal information into the estimation as well. Using HMM, they extract VQ-like clusters called *HMM states* and every speech frame is assigned to a state, which is associated with a wideband envelope. Again, the selection criteria among candidate states is the minimum euclidian distance.

Yannis *et al* [9] argue that narrowband envelopes have a one-to-many relationship with extension band envelopes. Utilizing this relationship, they propose an estimation/coding scheme. They test their schemes using two different envelope mapping techniques, known as *Non Linear Interpolative Vector Quantization (NLIVQ)* [10] and *GMM Conversion Function Estimator (GMMCF)* [7, 11]. To represent both narrowband and extension band envelopes, they use 10^{th} order Line Spectral Frequencies (LSFs).

Kontio *et al* [12] exploit neural networks to solve ABE problem and propose a method called *Neuroevolution Artificial Bandwidth Expansion (NEABE).* First, the method uses spectral folding to construct the initial spectral components at the extension band. Next, these components are shaped by a set of parameters which are decided by a neural network.

1.2 Contributions

In this thesis, we propose a novel method for the extension of the spectral envelope.

- Our method utilizes HMMs from narrowband to wideband envelope mapping.
- We introduce linear estimation into HMM mapping.
- We extract linear filters from narrowband envelope to wideband envelope based on HMM states.
- We investigate the effects of temporal neighborhood in linear estimation in addition to the contribution of HMM.
- We decrease the order of linear estimation by exploiting the correlation between the source narrowband feature vector to each instance of the desired wideband feature.

1.3 Outline

Here, we present an outline of the thesis. Chapter 2 briefly explains the Human Speech Reproduction System and Source-Filter separation. We also implement a benchmark system that uses Vector Quantization for mapping from narrowband envelope to wideband envelope. This benchmark system is described in this chapter as well.

Chapter 3 focuses on our proposed method. In this chapter, in order to define our model, we first explain Markov Models and Hidden Markov Models. We, then, reveal our model and explain the dynamics of the method in detail.

Chapter 4 declares the experimental results of the benchmark system and the proposed method. This chapter also includes the structure of our training and test databases, as well as a thorough discussion of the performance of the method with respect to varying parameters.

Finally, Chapter 5 includes a summary of the work presented and aims at possible future research objectives related to the subject and our contributions.

1.4 Notation

In this thesis, $(\cdot)^T$ corresponds to the transpose operation, $[\cdot]^{-1}$ represents Matrix inverse. Estimated scalars (or vectors) are represented as $\hat{x}$ (or $\hat{\boldsymbol{x}}$).

Chapter 2

SPEECH PRODUCTION SYSTEM AND BENCHMARK SYSTEM

2.1 Source-Filter Separation

In order to map narrowband sound signals to wideband signals, we need to have some a priori knowledge about our source signal. Since the scope of this thesis is artificial bandwidth extension of speech signal, we can easily argue that our source signal is speech signal, which is the first and by far the most important a priori knowledge that we have. Source-Filter separation is highly used in speech processing applications, and is an extremely useful model to represent speech signals [13, 14].

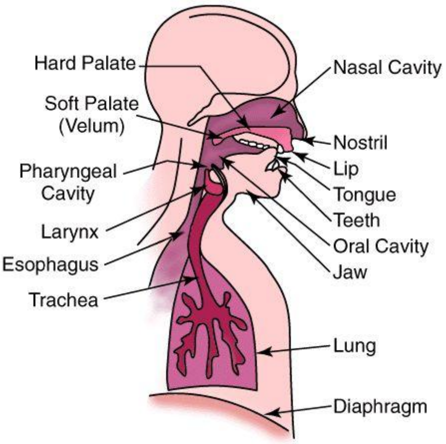


Figure 2.1: Voice Production System

Source-Filter separation takes its roots from the actual human voice production system. Fig 2.1 shows the anatomy of the human voice production system. The organs that take part in voice production are the lungs, trachea (windpipe), larynx (organ of voice production), pharyngal cavity (throat), oral cavity (mouth) and nasal cavity (nose).

The sound we perceive is the vibration of the air. The airflow that initiates this vibration comes from the lungs by the movement of the diaphragm. Throughout its path from the lungs to oral cavity, this airflow is given a characteristic pattern. The shape and position of the cavities and organs like tongue and teeth decides this pattern.

Source signal (excitation) is defined as the air that comes from the lungs and passes through vocal chords. The cavities and the other organs shape this airflow and produces the speech that we hear. Therefore, these cavities and organs behave like a filter that drives the source, which is the airflow. The name comes from this analogy.

2.1.1 Source Signal

As stated above, source signal is the airflow that comes from the lungs and passes through the vocal chords at the larynx. For a voiced sound, this airflow has a comb-like (pulse train) structure [13]. The period of this pulse train is decided by the tension and mass of the vocal chords. For unvoiced sounds, there is no tension on the vocal chords and the air passing through is significantly degraded, therefore the airflow has a noise like structure.

To demonstrate this acoustic source, Source-Filter model follows a similar pattern. Voiced sounds are represented as periodic pulse trains where this period is called *pitch frequency* and is typically between 125-250 Hz for women and 50-250 Hz for men. On the other hand, unvoiced sounds are represented as random noise, which is shown to be spectrally flat [3].

2.1.2 Filter

In Source-Filter model, the cavities and organs in the human voice production system define the shape and characteristics of the air that comes from the lungs through the vocal chords, therefore are together named as the filter.

A simple visualization of the voice production with Source-Filter model is given in Fig 2.2.

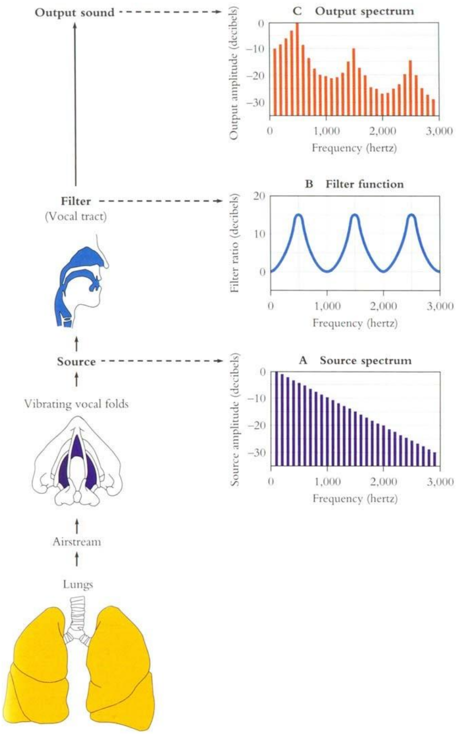


Figure 2.2: Excitation and Envelope Signals

2.2 Linear Predictive Coding

Studies show that [15] for a short time period (between ∼5ms and ∼100ms), speech signals behave as stationary signals. Therefore, speech signals may be represented as consecutive frames where each frame is a periodic stationary signal. Linear predictive coding states within these frames, speech signals within frames can be represented with an excitation signal and a finite order auto regressive filter and finds a solution to the problem of extracting this excitation and filter coefficients.

In discrete time signals, where we represent speech signals digitally, we define the source signal, filter function and the resulting speech signal as $E(z)$, $H(z)$ and $X(z)$, respectively. Therefore we can simply conclude that

$$X(z) = E(z)\,H(z) \tag{2.1}$$

If we define an inverse filter $A(z)$ such that

$$A(z) = 1/H(z) \tag{2.2}$$

we can simply conclude from (2.1) and (2.2) that

$$E(z) = X(z)\,A(z) \tag{2.3}$$

Assuming $A(z)$ is an all-pole filter with sufficient poles,

$$A(z) = 1 - \sum_{k=1}^{p} a[k]\,z^{-k} \tag{2.4}$$

In time domain, (2.3) and (2.4) correspond to

$$x[n] = \sum_{k=1}^{p} a[k]\,x[n-k] + e[n] \tag{2.5}$$

Since speech signals are short time stationary, using a source and filter, we can estimate the resulting signal as

$$\hat{x}[n] = \sum_{k=1}^{p} a[k]\,x[n-k] \tag{2.6}$$

and the estimation error, which is equal to the excitation signal is

$$e[n] = x[n] - \hat{x}[n] = x[n] - \sum_{k=1}^{p} a[k]\,x[n-k] \tag{2.7}$$

The solution to these equations which minimize $e[n]$ are $a[k]$ which are the Linear Prediction Coding Coefficients (LPC Coefficients) and are calculated using Levinson Durbin Recursion [16]. Here, p is the order of the estimation.

LPC coefficients are used to represent the filter function in the Source-Filter Separation since $a[k]$ are used to calculate $A(z)$, which is our filter function.

2.2.1 Line Spectrum Frequencies

Line Spectral Frequencies (LSF) are extracted from LPC coefficients. LSFs are a set of monotonically increasing numbers between 0 and 0.5 and can be calculated with a direct one-to-one mapping from LPC coefficients. Studies show that [17] they are good representatives of the spectral envelope and during this thesis, are used as the feature vectors for spectral envelope.

Line Spectral Frequencies (LSF) and Line Spectrum Pairs (LSP) are identical and may be used interchangeably during this thesis.

2.3 Benchmark VQ System

In order to be able to compare our proposed system, we first implement a benchmark system that utilizes Vector Quantization. In this system, we exploit the Source-Filter model, therefore our task is divided into two:

1. Extension of the excitation

2. Extension of the spectral envelope

Following sections explain both of these procedures.

2.3.1 Extension of the excitation

As the Source-Filter model states, the excitation signal has a comb-like structure for voiced sounds and is similar to white gaussian noise for unvoiced sounds. Furthermore, [4] states that the extension of the excitation plays a minor role compared to the extension of the envelope in terms of improving the quality of the perceived speech. Consequently, we employ a simple, yet effective spectral folding method to extend the excitation signal, which is given in (2.8);

$$\boldsymbol{e}^{wb}[k'] = \begin{cases} \boldsymbol{e}^{nb}[k] & \text{if } k' = 2k \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

2.3.2 Extension of the envelope

In order to map narrowband envelope to wideband envelope, vector quantization is a widely used technique, also employed by [5, 10].

For a given large set of data, vector quantization finds a small number of data points, each of which have approximately same number of points in their neighborhood. These points are called *centroids.* In other words, vector quantization manages to find a certain number of centroids that best define the large data set, since every instance in the data set is assigned to the nearest centroid with respect to a distance metric.

To exploit vector quantization for artificial bandwidth extension, we first have to decide on our data points, that is, our vectors. In section 4.1 we have stated that a parallel corpus with both narrowband and wideband versions of TIMIT database is obtained. This means, for every 20ms frame in a speech signal within the training database, LSF features that represent the spectral envelope are extracted as

$$\boldsymbol{f}_m^{nb} = [f_1^{nb} f_2^{nb} \dots f_{10}^{nb}]$$

and

$$\boldsymbol{f}_m^{wb} = [f_1^{wb} f_2^{wb} \dots f_{16}^{wb}]$$

where $\boldsymbol{f}_m^{nb}$ is the narrowband LSF vector at frame m and $\boldsymbol{f}_m^{wb}$ is the wideband LSF vector at frame m.

We, then, define our new feature vectors as the fused versions of these narrowband and wideband LSFs. For a given frame m, the new feature vector can be described as

$$\boldsymbol{f}_m = [\boldsymbol{f}_m^{nb} \boldsymbol{f}_m^{wb}]$$

which means

$$\boldsymbol{f}_m = [f_1^{nb} f_2^{nb} \dots f_{10}^{nb} f_1^{wb} f_2^{wb} \dots f_{16}^{wb}]$$

We have now constructed new feature vectors with dimension $10 + 16 = 26$. Using these new feature vectors, we train our model with VQ. This is done by utilizing LBG algorithm [6] which calculates C centroids that best define the data set and store these centroids in a codebook.

When the codebooks are extracted, training part is complete and we continue with the quantization part. For every narrowband frame, we calculate the distance between its feature vector and the first 10 components of every centroid in the codebook. The nearest codebook entry is decided as the candidate with the smallest euclidian distance to the source vector. This distance is defined as

$$\boldsymbol{x}^c - \boldsymbol{f}_m = \sqrt{\sum_{i=1}^{10} (x_i^c - f_i)^2} \tag{2.9}$$

where $\boldsymbol{x}^c$ is the c^{th} codebook entry.

When the nearest codebook entry is assigned, its wideband part, that is $[x_{11}x_{12}\ldots x_{26}]$, is decided as the wideband correspondent of the narrowband source vector, therefore mapping from narrowband to wideband is complete.

2.3.3 Summary of VQ Algorithm

i. Test signals are decomposed into excitation and envelope

ii. Excitation is extended with spectral folding

iii. For the training data, narrowband LSFs and wideband LSFs are grouped together to form fused $\boldsymbol{f}_m$ vectors

iv. Using these $\boldsymbol{f}_m$ vectors, VQ is trained and codebooks are extracted

v. Every narrowband envelope in the test data is compared to the first 10 instances of codebook entries and the nearest candidate is selected

vi. The wideband counterpart of the nearest codebook entry is assigned as the new estimated wideband envelope

vii. Using the estimated wideband envelope and the estimated wideband excitation, we construct the estimated wideband speech file

Chapter 3

PROPOSED MODEL FOR ARTIFICIAL BANDWIDTH EXTENSION OF THE ENVELOPE

In this chapter, we propose our new method that deals with the extension of the narrowband spectral envelope. The extension of the envelope incorporates linear estimation to Hidden Markov Models (HMMs). In order to explain our system effectively, we focus on Markov Models and Hidden Markov Models.

3.1 Markov Models

Markov models are probabilistic models widely used to define processes [15]. The definition on Markov models starts by considering a system with N states. In other words, at any given time, this system is in one of these N states $[s_1, s_2, \ldots, s_N]$.

Periodically, this system goes into a state transition where it changes its state to one of the N states $[s_1, s_2, \ldots, s_N]$, including the state it was already in. For simplicity, we will refer to the state of the system at time t as q_t.

For first order Markov models, the probability of system being at state q_t at time t is only dependent on the previous state q_{t-1}. Theoretically,

$$P\left[q_{t+1} = j | q_t = i\right] = P\left[q_{t+1} = j | q_t = i, q_{t-1} = k, \ldots, q_1 = n\right] \tag{3.1}$$

This probability, with which the system will change state from s_i to s_j is called *state transition probability from i to j* and is denoted as a_{ij}.

$$a_{ij} = P\left[q_{t+1} = j | q_t = i\right] \tag{3.2}$$

Since we know that there are N states, we can easily conclude that

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{3.3}$$

$$a_{ij} \geq 0 \quad \forall i, j \tag{3.4}$$

Furthermore, we define the system's probability to be at state i at any given time as s_i. This means

$$\sum_{i=1}^{N} s_i = 1 \tag{3.5}$$

$$s_i \geq 0 \quad \forall i \tag{3.6}$$

To sum up, the parameters that we need to define a Markov process with N states are s_i and a_{ij} where $1 \leq i, j \leq N$. A simple Markov model with three states is depicted in Fig. 3.1.

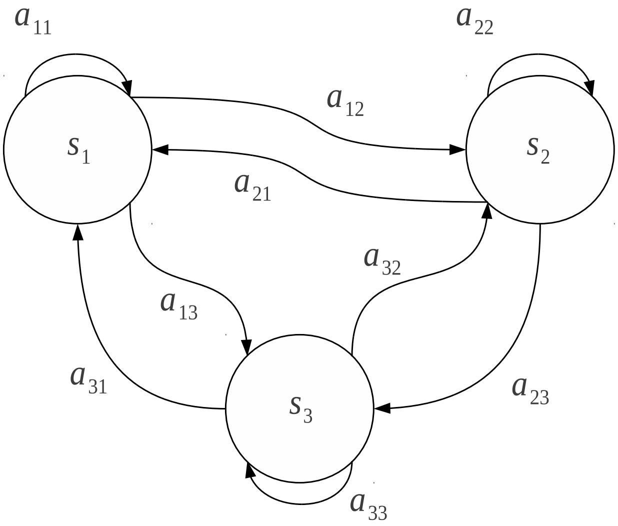


Figure 3.1: A Markov chain with 3 states and corresponding state transition probabilities

3.2 Hidden Markov Models

In Markov models, each state corresponds to a deterministic observable event, therefore the output of the system which can be observed is not a random variable. This makes Markov models insufficient to define real world processes.

Hidden Markov Models overcome this insufficiency of Markov Models. In HMMs, the observations actually are a probabilistic function of the state [15]. In other words, the actual model is a Markov Model with a finite number of states and the observations are a function of the states of this very model. This is better visualized in Fig. 3.2

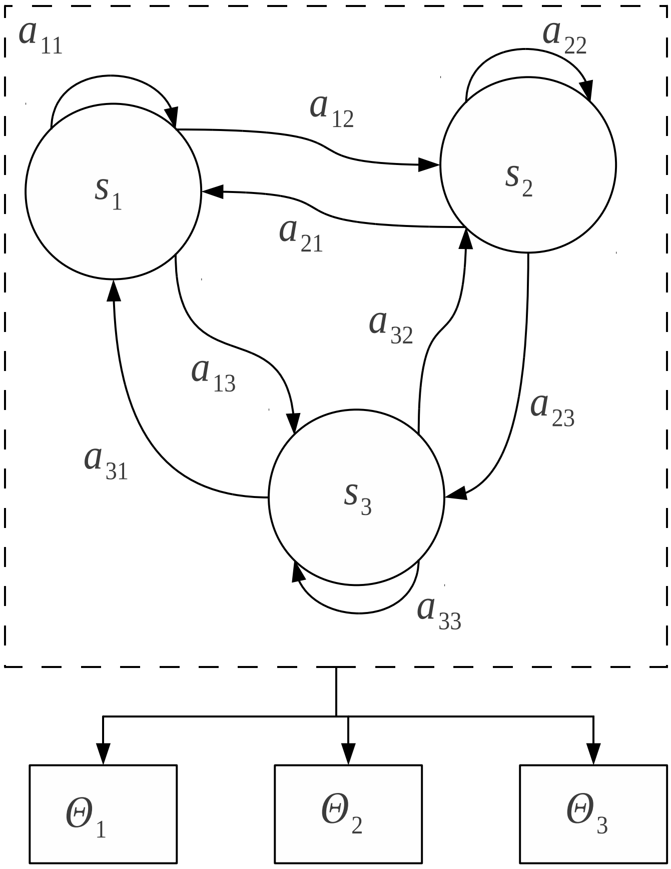


Figure 3.2: A Hidden Markov process with 3 states s_i, state transition probabilities a_{ij} and observable outputs Θ_i

3.2.1 Hidden Markov Model Parameters

The formal definition of HMMs are given in section 3.2. We now define the parameters needed to express a Hidden Markov Model thoroughly.

1. *The number of states (N):* The first parameter that defines a Hidden Markov Model is the number of states. It should be noted again that the states are not identical to the observations.

2. *The number of observation symbols:* In order to characterize a Hidden Markov Model, we also need to know the number of observation symbols per state in addition to the number of states N.

3. *State transition probabilities:* Similar to Markov chains, we need to know the state transition probabilities (a_{ij}) in order to define a Hidden Markov Model. Keeping in mind that transition from any state i state j is allowed, state transition probabilities are stored as a $N \times N$ matrix A where the element at the i^{th} row and j^{th} column A_{ij} is equal to a_{ij}.

$$A = \begin{bmatrix} a_{11} & \cdots & \cdots & \cdots & a_{1N} \\ \vdots & \ddots & & & \vdots \\ \vdots & & a_{ij} & & \vdots \\ \vdots & & & \ddots & \vdots \\ a_{N1} & \cdots & \cdots & \cdots & a_{NN} \end{bmatrix}$$

4. *Observation state probability distribution:* We have stated that the observations are a probabilistic function of the states, thus we need to explore the nature of this function in order to properly define an HMM.

5. *The initial state distribution (π):* We need to know the initial state distribution π for all $1 \leq i \leq N$ where

$$\pi_i = P\left[q_1 = i\right]$$

3.3 Proposed Model

For a given speech signal, we know that the narrowband spectral envelope and wideband spectral envelope are correlated. In order to explain this correlation we utilize Hidden Markov Models. The superiority of HMM over VQ and GMM is that the model favors temporal information, in other words, the state at a given time t is related to the previous state, which is a perfect model to simulate slowly varying speech signals.

We model the spectral envelopes with Line Spectrum Frequency (LSF) representation of linear prediction filter. From now on, we will refer to the narrowband features as

$$\boldsymbol{f}^n = [f_1^n, f_2^n, \ldots, f_{10}^n]$$

and the wideband features as

$$\boldsymbol{f}^w = [f_1^w, f_2^w, \ldots, f_{16}^w]$$

In temporal clustering, the LSF features together with their first and second derivatives are used and they are referred as

$$\boldsymbol{F} = [\boldsymbol{f}, \Delta\boldsymbol{f}, \Delta\Delta\boldsymbol{f}]$$

where

$$\boldsymbol{f} = [\boldsymbol{f}^n \boldsymbol{f}^w]$$

3.3.1 Temporal Clustering

In our model [18], we use an HMM based unsupervised multi-stream analysis framework to build a correlation model between the narrowband spectral envelope and the wideband spectral envelope [19,20]. This multi-stream analysis allows us to capture temporal clusters which are actually the recurring phonetic segments.

We train the multi-stream HMM structure Λ^{nw} with the narrowband and wideband joint feature stream $\boldsymbol{F}$. Here, the HMM structure Λ^{nw} is used for unsupervised temporal clustering. Λ^{nw} is composed of B parallel HMMs, $\{\lambda_1^{nw}, \lambda_2^{nw}, \ldots, \lambda_B^{nw}\}$, where each λ_b^{nw} is choosen to be single state HMM, $\{s_b\}$, as shown in Fig. 3.3.

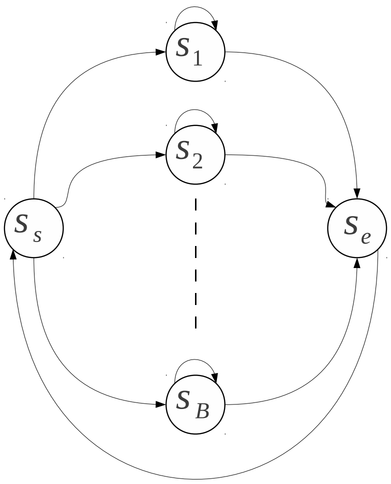


Figure 3.3: Parallel HMM structure, where s_s and s_e are non-emitting states.

Using this HMM model, we can decide on the states. Given the multimodal feature sequence, $\boldsymbol{F}^{nw} = \{\boldsymbol{F}_1^{nw}, \boldsymbol{F}_2^{nw}, \ldots, \boldsymbol{F}_K^{nw}\}$, $\boldsymbol{F}_k^{nw}$ denotes the joint feature vector at frame k. We then perform Viterbi decoding to find the state sequence $\boldsymbol{q}^{nw} = \{q_1^{nw}, q_2^{nw}, \ldots, q_K^{nw}\}$ which maximizes the probability of model match. The relation between the multimodal feature sequence $\boldsymbol{F}^{nw} = \{\boldsymbol{F}_1^{nw}, \boldsymbol{F}_2^{nw}, \ldots, \boldsymbol{F}_K^{nw}\}$, $\boldsymbol{F}_k^{nw}$ and the state sequence $\boldsymbol{q}^{nw} = \{q_1^{nw}, q_2^{nw}, \ldots, q_K^{nw}\}$ forms VQ-like clusters which also exhibit temporal information thanks to the nature of Hidden Markov Models.

3.3.2 Linear Estimation of the Wideband Spectra

We execute temporal clustering with the HMM model. Next, for each state, we extract linear filters to calculate the l^{th} wideband LSF feature from the highly correlated temporal and spatial neighborhoods of the narrowband LSF feature. For a given state s_b in Λ^{nw}, we define the mean removed narrowband LSF features at frame k as $\bar{\boldsymbol{f}}_k^n$ and the wideband mean removed LSF features at frame k as $\bar{\boldsymbol{f}}_k^w$.

We then define our source vector for the linear estimation part as the narrowband LSF features at the T^{th} temporal neighborhood, which can be denoted as

$$\boldsymbol{x} = [\bar{\boldsymbol{f}}^n_{k-T}, \ldots, \bar{\boldsymbol{f}}^n_k, \ldots, \bar{\boldsymbol{f}}^n_{k+T}]$$

Now that we have defined our source vector, we can state a linear estimator of the l^{th} mean removed feature of the wideband LSF $y = \bar{f}^w_{k,l}$ as

$$\hat{y} = \boldsymbol{x}\omega_{\mathbf{l}}^{\mathbf{T}} \tag{3.7}$$

where row vector $\omega_{\mathbf{l}}$ represents the $10(2T+1)^{th}$ order linear prediction filter for the l^{th} wideband LSF component and $[.]^T$ is the vector transpose operator. The linear estimator that gives the minimum estimation error between y and $\hat{y}$ can be calculated using Yule-Walker equations,

$$\boldsymbol{R}_{yx} = \boldsymbol{R}_{xx}\omega_{\mathbf{l}}^{\mathbf{T}} \tag{3.8}$$

where R_{yx_i} and $R_{x_ix_j}$ are the correlation of y, x_i and x_i, x_j signals, respectively $R_{yx_i} = E\{yx_i\}$ and $R_{x_ix_j} = E\{x_ix_j\}$. Given these equations, we can calculate $\omega_{\mathbf{l}}$ as

$$\omega_{\mathbf{l}}^{\mathbf{T}} = \boldsymbol{R}_{\mathbf{xx}}^{-\mathbf{1}}\boldsymbol{R}_{\mathbf{yx}} \tag{3.9}$$

where $[.]^{-1}$ is the matrix inverse operator.

3.3.3 Feature Selection

Using our model, we define $10(2T+1)^{th}$ order linear estimators to find l^{th} component of the wideband LSF feature vector. However, the source vector $\boldsymbol{x}$ is high dimensional and the LSF feature components are mostly correlated; therefore we introduce feature selection to our estimation.

Instead of $\boldsymbol{x} = [\bar{\boldsymbol{f}}^n_{k-T}, \ldots, \bar{\boldsymbol{f}}^n_k, \ldots, \bar{\boldsymbol{f}}^n_{k+T}]$, we define $\boldsymbol{x}' = [x_{i_1}, x_{i_2}, \ldots, x_{i_p}]$, such that $\{i_1, \ldots, i_p\}$ are the indexes of the largest p correlations in R_{yx}. This requires saving the indexes for the most correlated p for every target l^{th} wideband LSF feature and for every state s_b. Again, using Yule-Walker equations, we calculate $\omega_{\mathbf{l}}$. Finally, we can calculate the l^{th} wideband LSF feature as

$$\hat{f}^w_{k,l} = \boldsymbol{x}\omega_{\mathbf{l}}^{\mathbf{T}} + \mu^{\mathbf{w}}_{\mathbf{b,l}} \tag{3.10}$$

where $\mu^w_{b,l}$ is the l-th mean wideband feature component of state s_b. Using feature selection, we decreased the order of our linear estimator to p. Note that the p^{th} order LP filter ω is extracted for each feature component l in each state s_b.

3.3.4 Reducing Number of Filters

Our model requires saving p^{th} order filters and p indexes for every 16 wideband feature and every B state. In order to decrease this memory overhead, instead of saving p indexes for every l^{th} feature of B states, we calculate the mean of $\boldsymbol{R}_{yx}$ matrices over all B states, find the most correlated p instances, and use these indexes for all B states during estimation.

3.4 Summary of the Proposed Method

Our model is simply depicted in Fig. 3.4, where the upper left block displays the temporal clustering and the upper right block shows linear estimation. The multi-stream parallel branch HMM model Λ^{nw} that we obtain temporal clustering is split into two, which are the narrowband model Λ^n and the wideband model Λ^w. These two models have the same state transition probabilities and they have split observation probability density functions representing $\boldsymbol{f}^n$ and $\boldsymbol{f}^w$ features. The observation probability density functions can decidedly be divided into two for Gaussian densities with diagonal covariance.

In short, given the narrowband model Λ^n, state dependent linear estimators and narrowband speech, the flow of the wideband spectra estimation is described as follows:

i. The narrowband feature sequence, $\boldsymbol{F}^n$, is extracted from the narrowband speech.

ii. Temporal segmentation of the narrowband feature sequence $\boldsymbol{F}^n$ is performed using the HMM model Λ^n to extract temporal sub-phone patterns with a state sequence $\boldsymbol{q}^n$.

iii. The L linear predictors in state q^n_k are used to extract the acoustic feature estimate $\hat{\boldsymbol{f}}^w_k$ as described in (3.10).

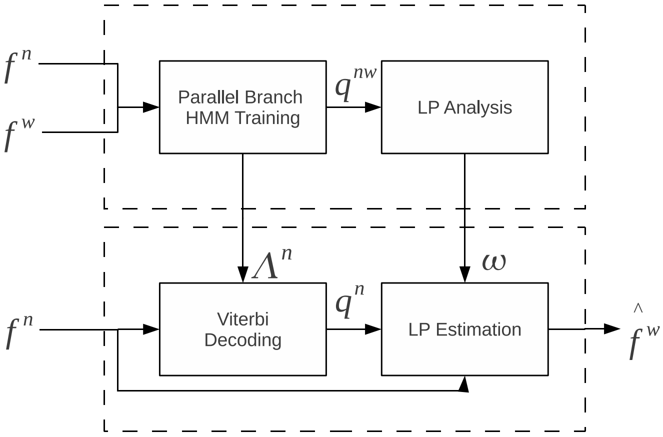


Figure 3.4: Proposed envelope extension method.

Chapter 4

EXPERIMENTAL RESULTS

4.1 The TIMIT Database

In this work, we have used a portion of the TIMIT database [21] which is a corpus of speech utterances by speakers from 8 different dialect regions of America. TIMIT is designed by a joint collaboration of Texas Instruments, Massachusetts Institute of Technology and Stanford Research Institute.

Specifically, we have used two different training and test sets during the experimentation. The first set was composed of *sa* sentences, which are dialect sentences that aim to expose the dialectical variance among speakers.

The second set was composed of *sx* sentences, which are phonetically compact sentences that were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest.

In both cases, the training set and the test set do not contain any similar file, in other words, they are exclusive.

Table 4.1: The Structure of Our TIMIT Database.

	Sentences (Speakers)	
	sa	*sx*
Training Set	924 (462)	1155 (231)
Test Set	336 (168)	420 (84)

The TIMIT database is composed of sentences sampled with 16kHz, in other words, wideband signals. However, both during the training and test phases, we will need and use narrowband signals as well. Therefore, we downsample all speech signals by 2 and obtain narrowband versions. That is, we accomplish to have a parallel corpus of the TIMIT database, with narrowband and wideband versions. In both the narrowband and wideband cases, the speech signals in this work are analyzed over 20ms frames.

In order to represent a narrowband envelope frame, we use 10^{th} order LSFs. Therefore, the narrowband envelope of the m^{th} frame is represented as

$$\boldsymbol{f}_m^{nb} = [f_1^{nb} f_2^{nb} \dots f_{10}^{nb}]$$

Similarly, we use 16^{th} order LSFs to represent a wideband envelope frame. Therefore, the wideband envelope of the m^{th} frame is

$$\boldsymbol{f}_m^{wb} = [f_1^{wb} f_2^{wb} \dots f_{16}^{wb}]$$

4.2 Objective Metrics

4.2.1 Segmental Signal-to-Noise Ratio

For speech signals, instead of traditional SNR, a frame based distance metric, namely *Segmental SNR*, which is the mean of SNR values over frames is defined. The SegSNR distance between a target vector x_t and a source vector x_s is calculated as

$$SegSNR_{ts} = \frac{10}{M} \sum_{m=0}^{M-1} log \frac{\sum_{n=Nm}^{n=Nm+N-1} x_s^2(n)}{\sum_{n=Nm}^{n=Nm+N-1} [x_t(n) - x_s(n)]^2} \tag{4.1}$$

where M is the number of frames and N is the dimension of each frame.

4.2.2 Logarithmic Spectral Distortion

For a given frame m, squared logarithmic spectral distortion between a wideband envelope $A_w(e^{j\Omega}; m)$ and an estimated envelope $\hat{A}_w(e^{j\Omega}; m)$ can be calculated as

$$d_{LSD}^2 = \frac{1}{N} \sum_{k=0}^{N-1} (20log_{10} \frac{\sigma_w}{|A_w(e^{j\Omega}; m)|} - 20log_{10} \frac{\hat{\sigma}_w}{|\hat{A}_w(e^{j\Omega}; m)|})^2 \tag{4.2}$$

where N is the Discrete Fourier Transform order and σ_w and $\hat{\sigma}_w$ are gain values for the original and estimated wideband envelopes, respectively. For signals composed of M frames, root mean square (RMS) LSD value is given as

$$\hat{d}_{LSD} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} d_{LSD}^2(m)} \tag{4.3}$$

4.2.3 Perceptual Evaluation of Speech Quality

Perceptual Evaluation of Speech Quality (PESQ) is a ITU-T P.835 recommendation which is used to compare speech signals. Further information can be found in [22].

4.3 Performance of the Benchmark VQ System

The benchmark system that utilizes vector quantization is tested with our three objective distance metrics. The system is trained for different codebook sizes ranging from 2 bits to 8 bits. As explained in section 4.1, the system is trained and tested on both SA and SX sets. The results are given in Table 4.2 and Fig.4.1.

Table 4.2: Performance of the VQ based benchmark ABE system.

CB Size	SegSNR(dB)		LSD(dB)		PESQ	
	SA	SX	SA	SX	SA	SX
4	8.503	8.377	5.680	5.609	2.802	2.711
8	7.898	8.134	5.294	5.217	2.840	2.899
16	7.871	8.251	5.019	4.999	3.035	2.976
32	7.765	8.227	4.829	4.802	3.084	3.070
64	7.780	8.314	4.700	4.741	3.200	3.144
128	7.751	8.301	4.647	4.666	3.262	3.214
256	7.760	8.279	4.592	4.617	3.309	3.255

As expected in the theory, increasing the codebook size yields a better estimation for vector quantization. For LSD measure, we observe for both sets that the score drops from 5.6dB to 4.6dB as we increase the codebook size from 4 to 256. We see an improvement in terms of PESQ score as well. PESQ score rises from 2.8 to 3.2. The SegSNR scores do not represent such a linear improvement with increasing codebook size, however, this measure is not correlated with speech quality as the other two [23].

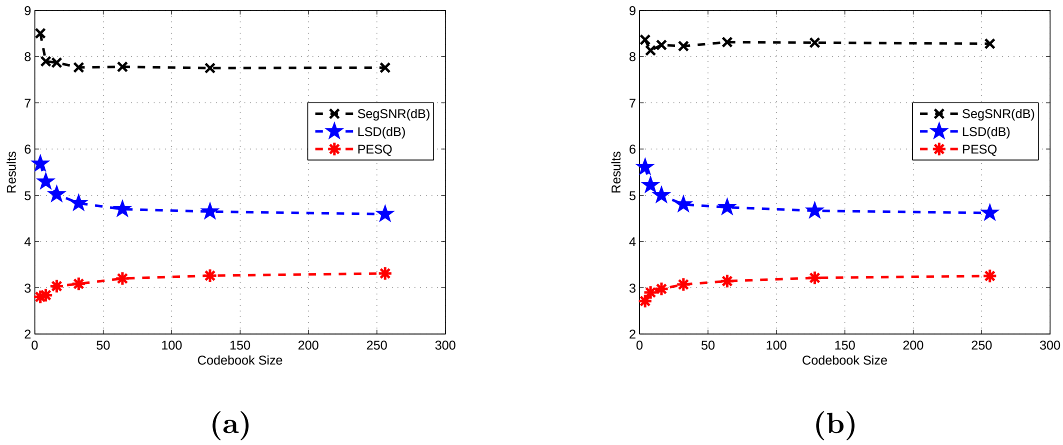


Figure 4.1: (a)Experimental Results for the SA set (b)Experimental Results for the SX set

4.4 Performance of the Proposed System

We implement and test our system as explained in Section 3.3 on both SA and SX sets. We use spectral folding for the extension of the excitation.

During the implementation, we change the branch number B from 4 to 256 by the order of 2 (which is identical to changing from 2 bits to 8 bits as in VQ case). The order of the linear estimation, p, is varied from 2 to 10. Additionally, the effect of the temporal neighborhood in linear estimation is observed for two different cases where $T = 0$ and $T = 1$. The effect of saving different indexes for every state s_b and using an average index for all cases is also examined for all cases given above.

We will present the objective results for each cases and later these results will be discussed. Note that the case with $T = 1$ and $p = 30$ corresponds to using a temporal neighborhood of 1 and not applying feature selection.

Table 4.3: SegSNR scores for SA set, $T = 0$, indexes are saved for every state s_b

SegSNR(dB)									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	7.705	8.092	7.746	7.648	7.497	7.239	7.209	7.118	7.085
8	8.651	8.226	8.142	8.163	8.023	7.802	7.628	7.552	7.509
16	7.850	7.763	7.667	7.574	7.449	7.504	7.418	7.346	7.312
32	8.196	8.177	8.234	8.169	7.997	7.901	7.813	7.728	7.664
64	7.951	7.915	7.856	7.831	7.788	7.668	7.576	7.528	7.502
128	8.109	7.979	7.954	7.924	7.838	7.718	7.643	7.591	7.553
256	8.105	8.043	8.025	7.977	7.892	7.781	7.688	7.645	7.603

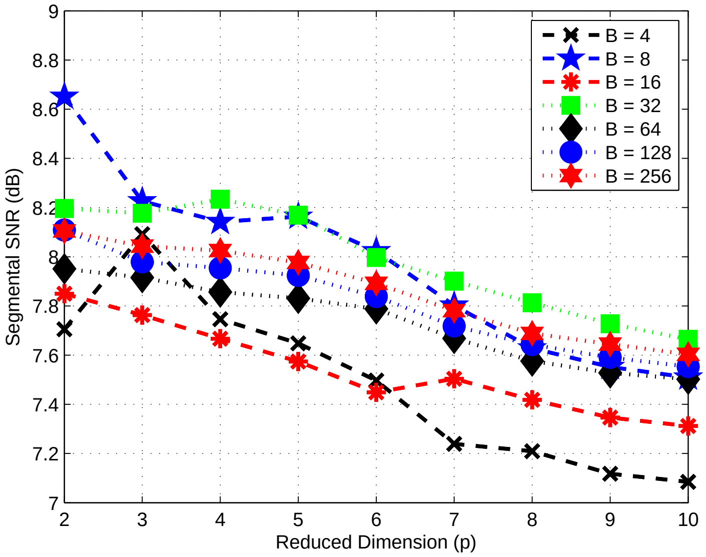


Figure 4.2: SegSNR scores for SA set, $T = 0$, indexes are saved for every state s_b

Table 4.4: SegSNR scores for SX set, $T = 0$, indexes are saved for every state s_b

SegSNR(dB)									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	10.351	9.932	9.348	9.371	9.158	9.011	8.908	8.916	8.831
8	8.681	8.720	8.612	8.593	8.449	8.360	8.249	8.235	8.182
16	8.654	8.561	8.650	8.515	8.431	8.366	8.258	8.176	8.164
32	9.787	9.685	9.531	9.509	9.360	9.275	9.060	8.916	8.858
64	9.038	8.896	8.885	8.826	8.721	8.619	8.540	8.466	8.405
128	9.456	9.439	9.363	9.216	9.136	9.043	8.919	8.850	8.770
256	9.932	9.020	8.995	8.910	8.777	8.704	8.610	8.510	8.449

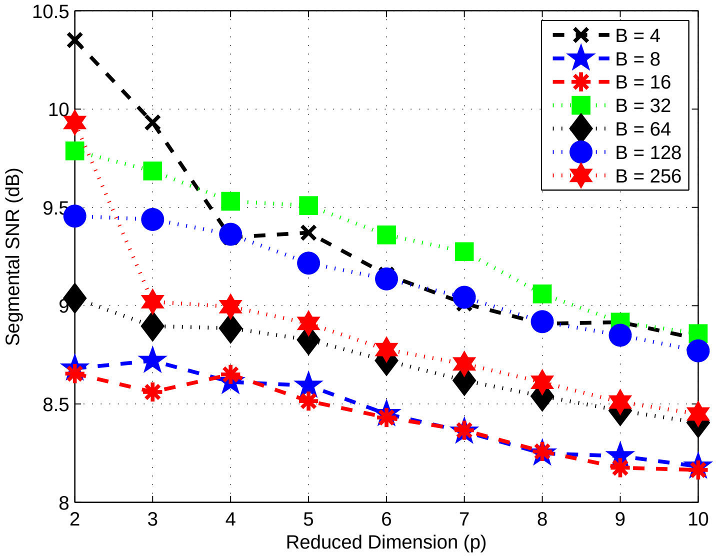


Figure 4.3: SegSNR scores for SX set, $T = 0$, indexes are saved for every state s_b

Table 4.5: LSD scores for SA set, $T = 0$, indexes are saved for every state s_b

LSD(dB)									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	4.323	4.174	4.113	4.042	4.026	3.973	3.951	3.931	3.918
8	4.027	3.905	3.848	3.812	3.775	3.760	3.744	3.727	3.716
16	3.815	3.715	3.679	3.650	3.629	3.609	3.599	3.587	3.576
32	3.781	3.673	3.633	3.597	3.584	3.570	3.559	3.548	3.537
64	3.644	3.536	3.505	3.487	3.473	3.460	3.449	3.441	3.433
128	3.548	3.462	3.434	3.417	3.404	3.392	3.383	3.374	3.367
256	3.514	3.430	3.401	3.385	3.371	3.361	3.353	3.346	3.341

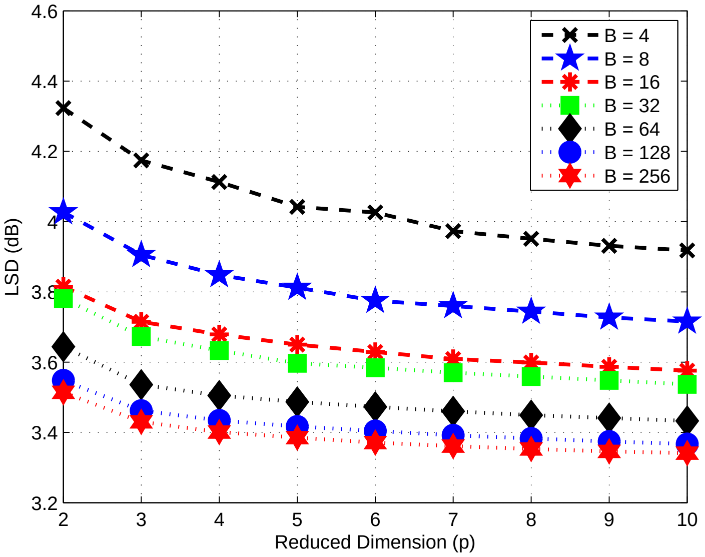


Figure 4.4: LSD scores for SA set, $T = 0$, indexes are saved for every state s_b

Table 4.6: LSD scores for SX set, $T = 0$, indexes are saved for every state s_b

LSD(dB)									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	4.218	4.088	4.040	4.016	3.983	3.967	3.939	3.930	3.916
8	4.024	3.904	3.835	3.814	3.789	3.748	3.740	3.725	3.712
16	3.791	3.695	3.659	3.628	3.609	3.594	3.580	3.567	3.560
32	3.757	3.666	3.618	3.592	3.569	3.558	3.546	3.534	3.523
64	3.614	3.514	3.483	3.465	3.447	3.438	3.431	3.422	3.411
128	3.503	3.422	3.394	3.376	3.364	3.354	3.346	3.337	3.331
256	3.427	3.353	3.327	3.311	3.299	3.291	3.283	3.276	3.271

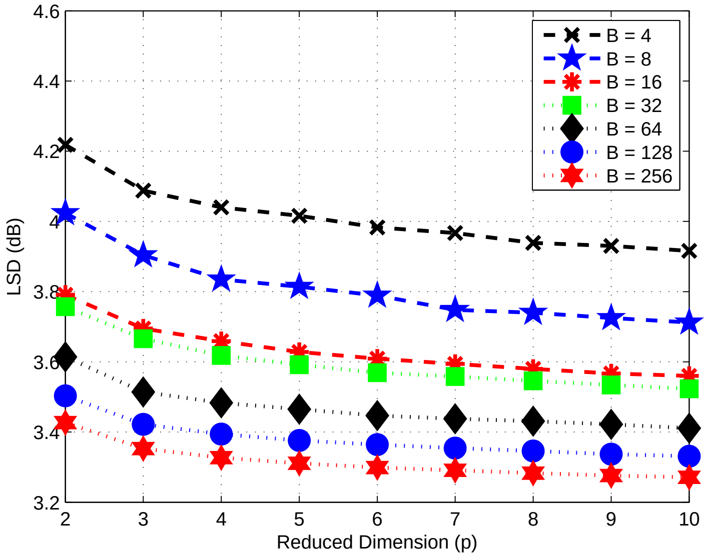


Figure 4.5: LSD scores for SX set, $T = 0$, indexes are saved for every state s_b

Table 4.7: PESQ scores for SA set, $T = 0$, indexes are saved for every state s_b

PESQ									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	3.585	3.826	3.826	3.859	3.858	3.843	3.855	3.845	3.843
8	3.875	3.938	3.941	3.943	3.952	3.944	3.943	3.941	3.941
16	3.897	3.923	3.925	3.930	3.931	3.929	3.933	3.933	3.933
32	3.849	3.909	3.928	3.940	3.943	3.947	3.946	3.950	3.951
64	3.794	3.869	3.878	3.880	3.883	3.887	3.888	3.889	3.891
128	3.853	3.896	3.909	3.910	3.915	3.919	3.922	3.924	3.924
256	3.870	3.901	3.915	3.920	3.924	3.925	3.924	3.927	3.927

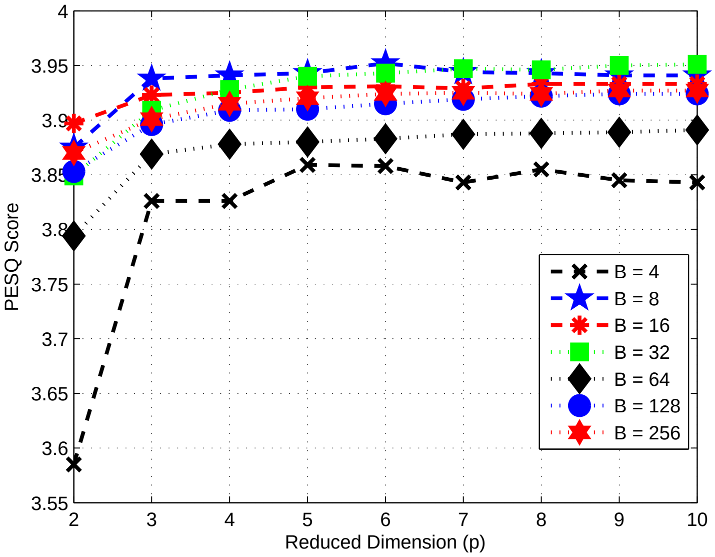


Figure 4.6: PESQ scores for SA set, $T = 0$, indexes are saved for every state s_b

Table 4.8: PESQ scores for SX set, $T = 0$, indexes are saved for every state s_b

PESQ									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	3.871	3.908	3.910	3.907	3.897	3.884	3.871	3.856	3.850
8	3.818	3.862	3.889	3.887	3.877	3.866	3.859	3.857	3.858
16	3.815	3.849	3.854	3.853	3.857	3.849	3.852	3.851	3.851
32	3.862	3.907	3.925	3.926	3.921	3.917	3.920	3.913	3.911
64	3.869	3.912	3.904	3.899	3.890	3.890	3.890	3.892	3.888
128	3.856	3.875	3.884	3.886	3.878	3.876	3.873	3.875	3.873
256	3.849	3.878	3.884	3.881	3.876	3.873	3.872	3.874	3.870

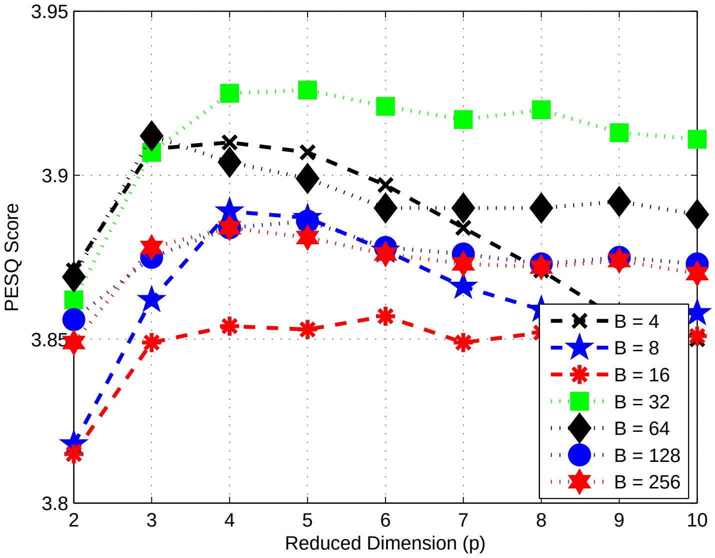


Figure 4.7: PESQ scores for SX set, $T = 0$, indexes are saved for every state s_b

Table 4.9: SegSNR scores for SA set, $T = 1$, indexes are saved for every state s_b

SegSNR(dB)										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	8.164	7.462	7.626	7.558	7.264	8.265	8.280	8.240	8.131	7.131
8	8.919	8.473	8.557	8.557	8.564	8.644	8.599	8.555	8.330	7.494
16	8.133	8.074	7.901	7.998	8.983	7.945	7.910	7.876	7.855	7.393
32	8.132	8.382	8.412	8.498	8.448	8.478	8.460	8.454	8.319	7.481
64	7.939	8.059	8.112	8.079	8.061	8.023	7.995	7.966	7.896	7.323
128	8.117	8.202	8.184	8.157	8.085	8.025	8.000	7.967	7.899	7.344
256	7.991	8.147	8.165	8.146	8.086	8.052	8.040	7.982	7.920	7.343

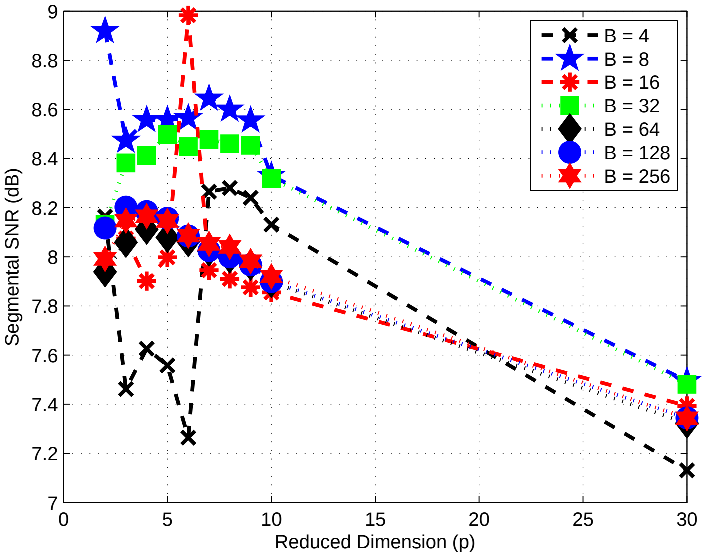


Figure 4.8: SegSNR scores for SA set, $T = 1$, indexes are saved for every state s_b

Table 4.10: SegSNR scores for SX set, $T = 1$, indexes are saved for every state s_b

SegSNR(dB)										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	10.465	11.031	10.718	10.866	10.579	10.328	10.040	10.032	9.587	8.847
8	9.155	9.233	9.030	9.108	9.068	8.823	8.879	8.813	8.826	8.281
16	9.074	9.132	8.830	8.896	8.853	8.866	8.865	8.832	8.700	8.247
32	10.014	10.162	10.116	10.123	10.081	10.148	10.034	9.928	9.716	8.683
64	9.078	9.180	9.114	9.160	9.147	9.170	9.138	9.046	8.991	8.314
128	9.434	9.584	9.669	9.687	9.651	9.623	9.577	9.459	9.339	8.525
256	9.094	9.142	9.209	9.179	9.110	9.116	9.085	9.026	8.999	8.280

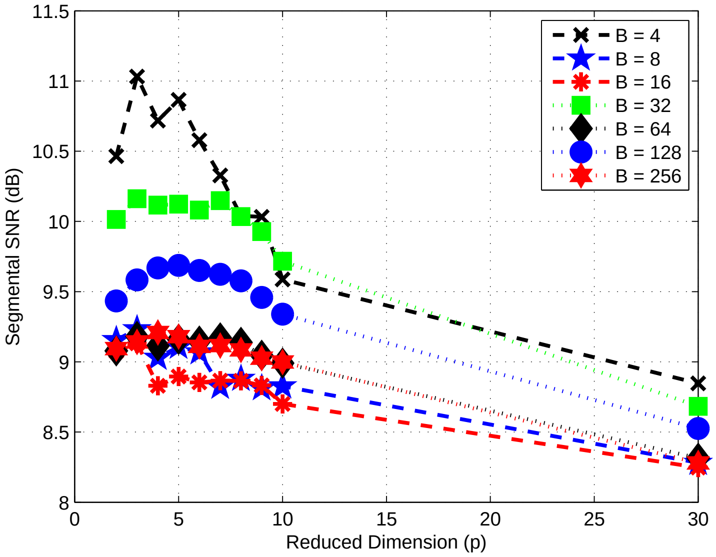


Figure 4.9: SegSNR scores for SX set, $T = 1$, indexes are saved for every state s_b

Table 4.11: LSD scores for SA set, $T = 1$, indexes are saved for every state s_b

LSD(dB)										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	4.353	4.271	4.206	4.188	4.167	4.087	4.055	4.046	4.025	3.859
8	4.096	3.997	3.911	3.903	3.883	3.839	3.831	3.816	3.787	3.652
16	3.860	3.795	3.730	3.718	3.699	3.669	3.659	3.649	3.630	3.515
32	3.842	3.754	3.691	3.670	3.644	3.619	3.608	3.598	3.578	3.468
64	3.690	3.611	3.554	3.527	3.513	3.491	3.481	3.472	3.458	3.377
128	3.597	3.521	3.467	3.457	3.435	3.417	3.409	3.401	3.390	3.317
256	3.562	3.484	3.436	3.414	3.397	3.382	3.373	3.367	3.358	3.305

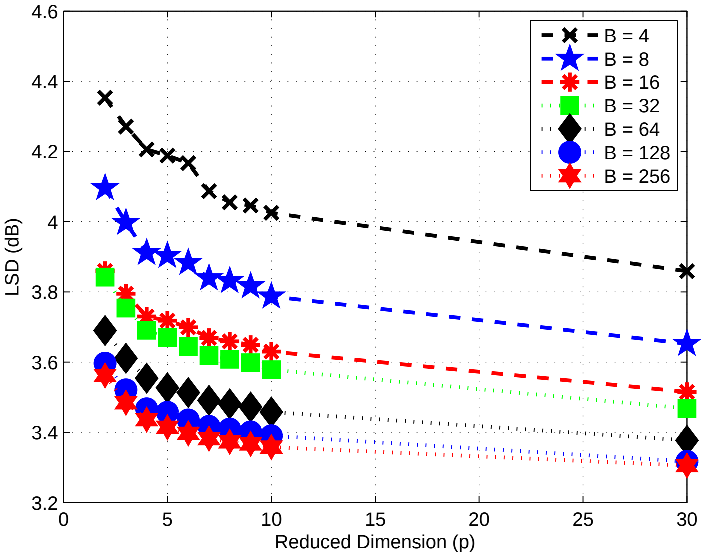


Figure 4.10: LSD scores for SA set, $T = 1$, indexes are saved for every state s_b

Table 4.12: LSD scores for SX set, $T = 1$, indexes are saved for every state s_b

LSD(dB)										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	4.283	4.211	4.143	4.109	4.102	4.034	4.004	4.000	3.971	3.843
8	4.077	4.001	3.936	3.896	3.873	3.834	3.819	3.810	3.795	3.643
16	3.843	3.773	3.713	3.692	3.678	3.648	3.635	3.628	3.607	3.504
32	3.807	3.731	3.676	3.657	3.642	3.617	3.605	3.594	3.574	3.464
64	3.659	3.584	3.527	3.511	3.499	3.479	3.470	3.462	3.449	3.365
128	3.547	3.477	3.429	3.412	3.398	3.382	3.375	3.366	3.357	3.286
256	3.469	3.404	3.360	3.348	3.355	3.318	3.311	3.304	3.294	3.244

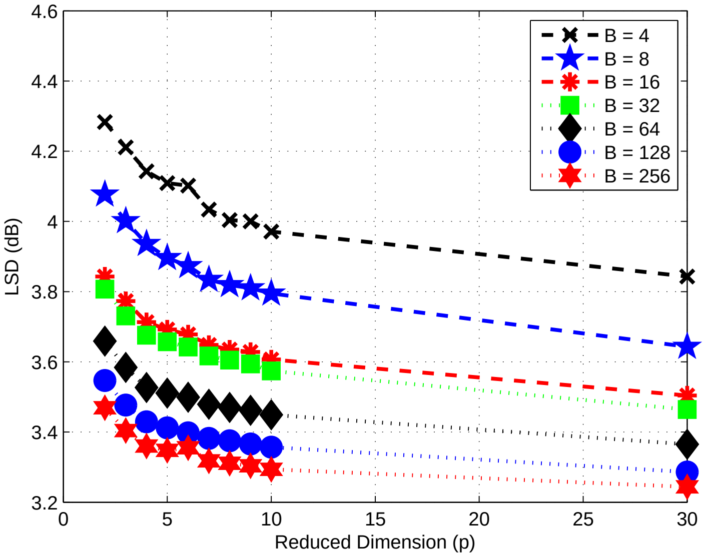


Figure 4.11: LSD scores for SX set, $T = 1$, indexes are saved for every state s_b

Table 4.13: PESQ scores for SA set, $T = 1$, indexes are saved for every state s_b

PESQ										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	3.488	3.563	3.586	3.586	3.591	3.827	3.855	3.857	3.860	3.875
8	3.821	3.873	3.914	3.916	3.936	3.926	3.931	3.940	3.944	3.952
16	3.839	3.887	3.894	3.895	3.896	3.908	3.915	3.918	3.920	3.950
32	3.709	3.802	3.853	3.869	3.873	3.911	3.920	3.924	3.929	3.966
64	3.747	3.806	3.795	3.850	3.825	3.867	3.874	3.877	3.883	3.908
128	3.748	3.785	3.849	3.830	3.886	3.895	3.898	3.906	3.911	3.928
256	3.750	3.794	3.855	3.867	3.881	3.887	3.900	3.902	3.914	3.902

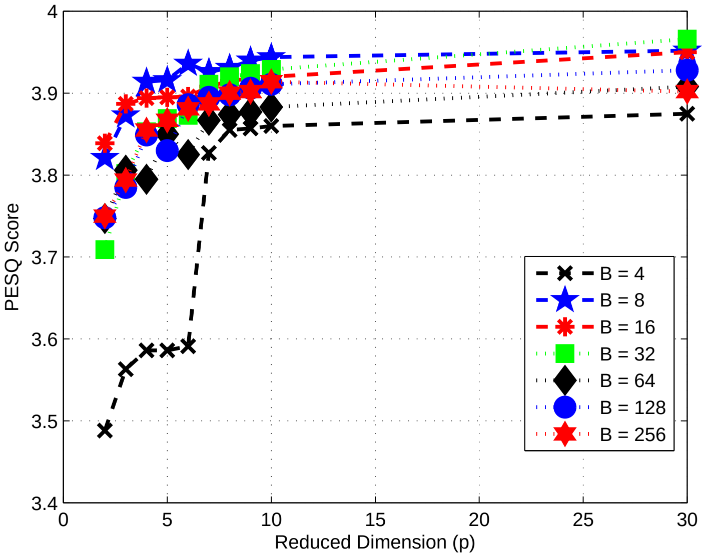


Figure 4.12: PESQ scores for SA set, $T = 1$, indexes are saved for every state s_b

Table 4.14: PESQ scores for SX set, $T = 1$, indexes are saved for every state s_b

PESQ										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	3.824	3.868	3.887	3.898	3.924	3.934	3.919	3.916	3.924	3.895
8	3.785	3.809	3.839	3.852	3.861	3.857	3.872	3.875	3.877	3.873
16	3.776	3.779	3.793	3.797	3.803	3.833	3.840	3.835	3.843	3.864
32	3.823	3.853	3.879	3.883	3.893	3.897	3.903	3.907	3.912	3.908
64	3.821	3.845	3.877	3.889	3.891	3.892	3.894	3.895	3.896	3.884
128	3.786	3.801	3.841	3.850	3.851	3.853	3.855	3.868	3.874	3.861
256	3.772	3.813	3.847	3.849	3.855	3.861	3.865	3.866	3.868	3.850

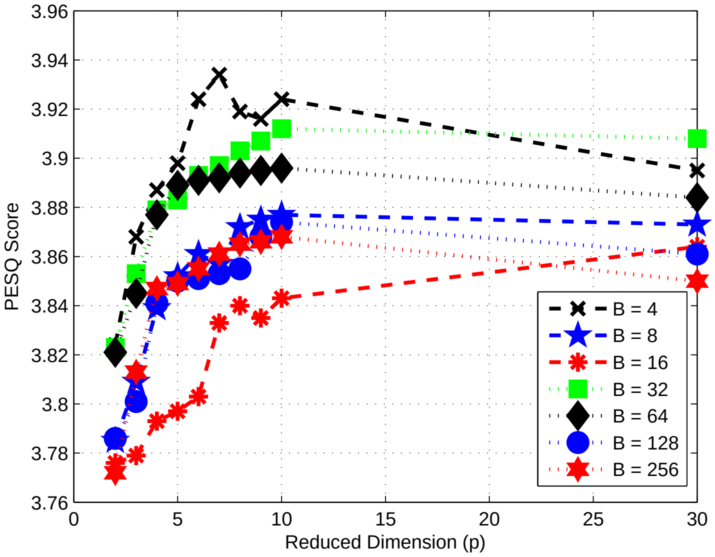


Figure 4.13: PESQ scores for SX set, $T = 1$, indexes are saved for every state s_b

Table 4.15: SegSNR scores for SA set, $T = 0$, single index saved

SegSNR(dB)									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	6.399	7.732	7.909	7.913	7.981	7.893	7.367	7.214	7.085
8	6.521	8.023	8.057	8.118	7.971	8.160	7.692	7.554	7.509
16	6.956	7.843	7.734	7.794	7.711	7.849	7.476	7.385	7.312
32	7.476	7.841	8.256	8.276	8.204	8.264	7.838	7.735	7.664
64	7.601	7.813	7.826	7.966	7.905	7.958	7.680	7.606	7.502
128	7.710	7.874	7.891	8.098	8.019	8.052	7.755	7.683	7.553
256	7.840	7.883	8.024	8.125	8.068	8.105	7.797	7.718	7.603

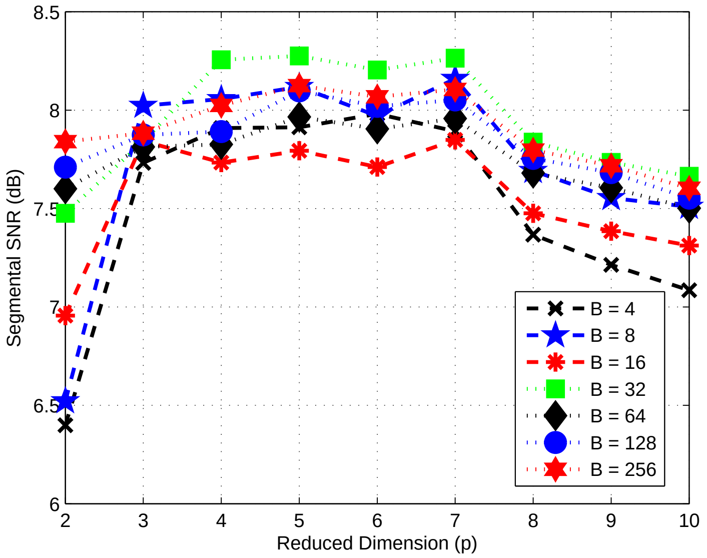


Figure 4.14: SegSNR scores for SA set, $T = 0$, single index saved

Table 4.16: SegSNR scores for SX set, $T = 0$, single index saved

SegSNR(dB)									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	6.936	8.099	8.295	8.790	9.286	8.555	8.558	8.517	8.831
8	6.910	7.750	7.770	8.248	8.566	8.028	7.904	7.895	8.182
16	7.386	8.034	7.931	8.149	8.357	7.684	7.924	7.918	8.164
32	7.670	8.343	8.464	8.812	9.117	8.516	8.596	8.541	8.858
64	8.030	8.386	8.240	8.413	8.646	8.177	8.229	8.219	8.405
128	8.200	8.523	8.496	8.646	8.945	8.464	8.563	8.532	8.770
256	8.326	8.498	8.374	8.476	8.742	8.296	8.304	8.286	8.449

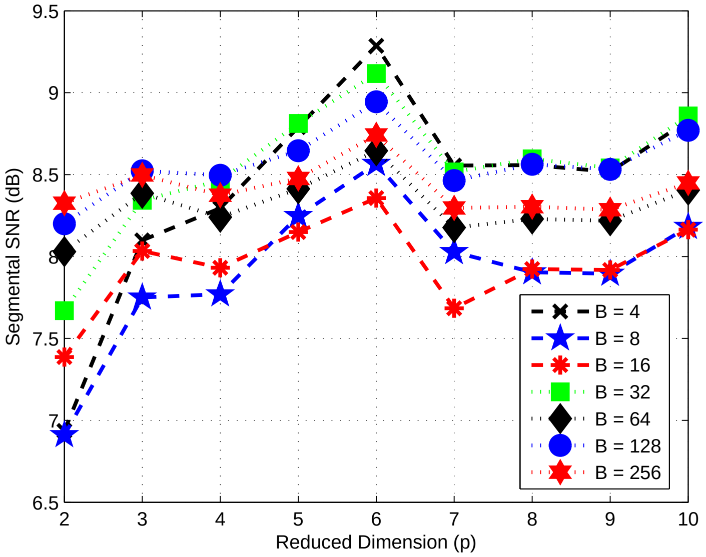


Figure 4.15: SegSNR scores for SX set, $T = 0$, single index saved

Table 4.17: LSD scores for SA set, $T = 0$, single index saved

LSD(dB)									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	5.438	4.786	4.436	4.358	4.293	4.164	4.059	4.019	3.918
8	5.049	4.544	4.216	4.139	4.075	3.963	3.853	3.824	3.716
16	4.718	4.311	4.033	3.957	3.906	3.808	3.709	3.683	3.576
32	4.559	4.231	3.955	3.893	3.847	3.755	3.666	3.636	3.537
64	4.284	4.022	3.893	3.750	3.708	3.629	3.549	3.525	3.433
128	4.151	3.920	3.794	3.660	3.623	3.549	3.477	3.456	3.367
256	4.075	3.891	3.736	3.609	3.577	3.508	3.441	3.423	3.341

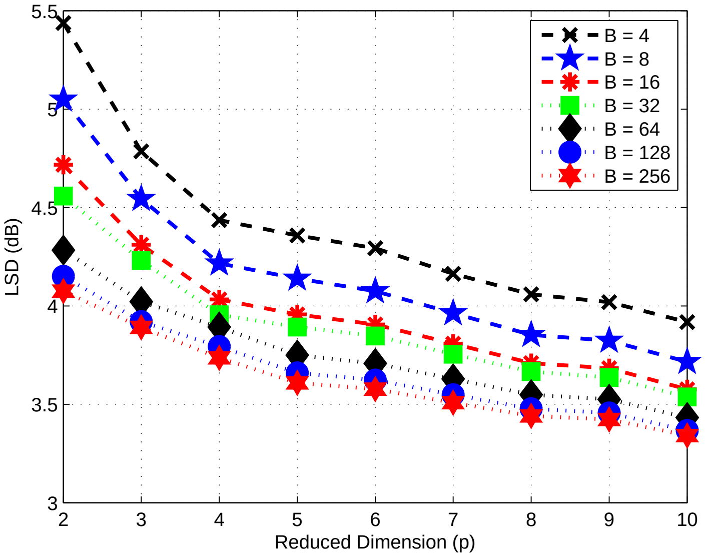


Figure 4.16: LSD scores for SA set, $T = 0$, single index saved

Table 4.18: LSD scores for SX set, $T = 0$, single index saved

LSD(dB)									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	5.166	4.655	4.415	4.307	4.227	4.129	4.046	4.026	3.916
8	4.912	4.423	4.185	4.087	4.008	3.915	3.828	3.803	3.712
16	4.547	4.180	3.975	3.904	3.842	3.753	3.672	3.646	3.560
32	4.468	4.123	3.923	3.854	3.788	3.712	3.632	3.608	3.523
64	4.229	3.949	3.766	3.714	3.656	3.588	3.516	3.493	3.411
128	4.051	3.814	3.647	3.605	3.553	3.494	3.429	3.410	3.331
256	3.939	3.721	3.568	3.531	3.482	3.428	3.364	3.346	3.271

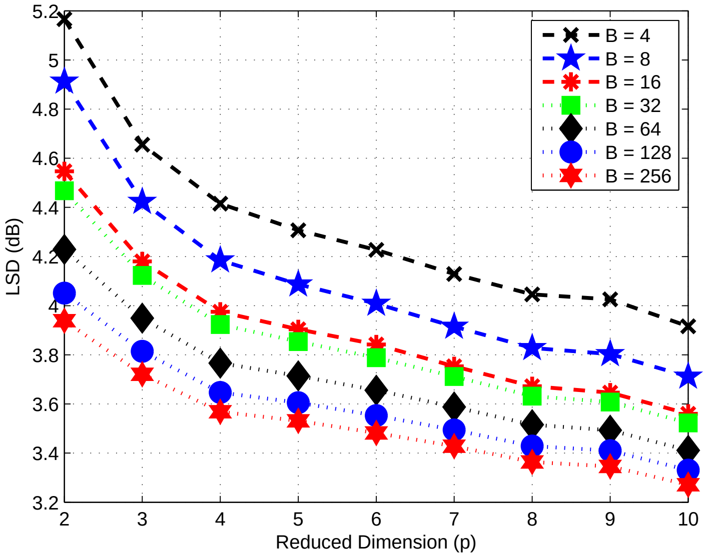


Figure 4.17: LSD scores for SX set, $T = 0$, single index saved

Table 4.19: PESQ scores for SA set, $T = 0$, single index saved

PESQ									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	2.656	2.895	3.608	3.720	3.694	3.751	3.819	3.799	3.843
8	2.798	3.134	3.698	3.789	3.759	3.963	3.898	3.894	3.941
16	2.881	3.238	3.690	3.780	3.759	3.821	3.891	3.892	3.933
32	3.037	3.223	3.750	3.820	3.822	3.822	3.913	3.908	3.951
64	3.212	3.502	3.641	3.750	3.792	3.806	3.863	3.867	3.891
128	3.232	3.529	3.671	3.660	3.840	3.856	3.903	3.902	3.924
256	3.307	3.504	3.715	3.857	3.857	3.875	3.907	3.910	3.927

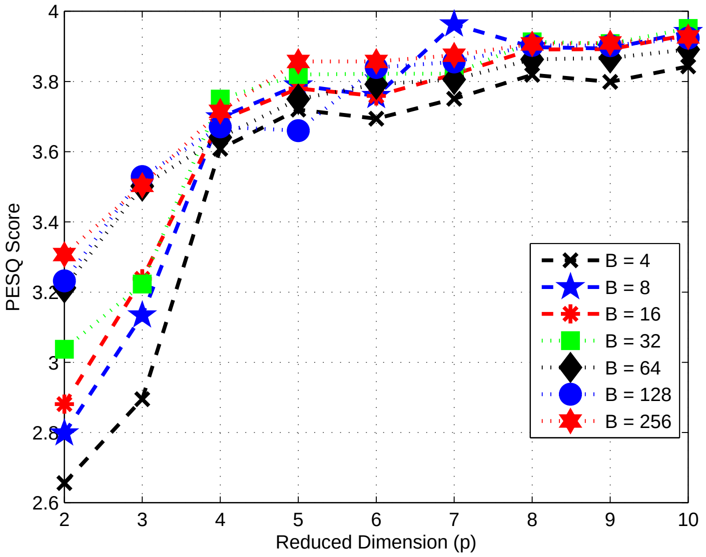


Figure 4.18: PESQ scores for SA set, $T = 0$, single index saved

Table 4.20: PESQ scores for SX set, $T = 0$, single index saved

PESQ									
B	Order of the estimation p								
	2	3	4	5	6	7	8	9	10
4	2.835	3.172	3.550	3.754	3.810	3.744	3.811	3.803	3.850
8	2.854	3.259	3.565	3.748	3.804	3.745	3.798	3.791	3.858
16	2.961	3.347	3.616	3.727	3.794	3.748	3.801	3.800	3.851
32	3.031	3.329	3.615	3.796	3.869	3.811	3.857	3.852	3.911
64	3.181	3.473	3.690	3.787	3.842	3.809	3.857	3.853	3.888
128	3.261	3.528	3.724	3.781	3.833	3.800	3.842	3.841	3.873
256	3.335	3.580	3.750	3.787	3.833	3.804	3.842	3.842	3.870

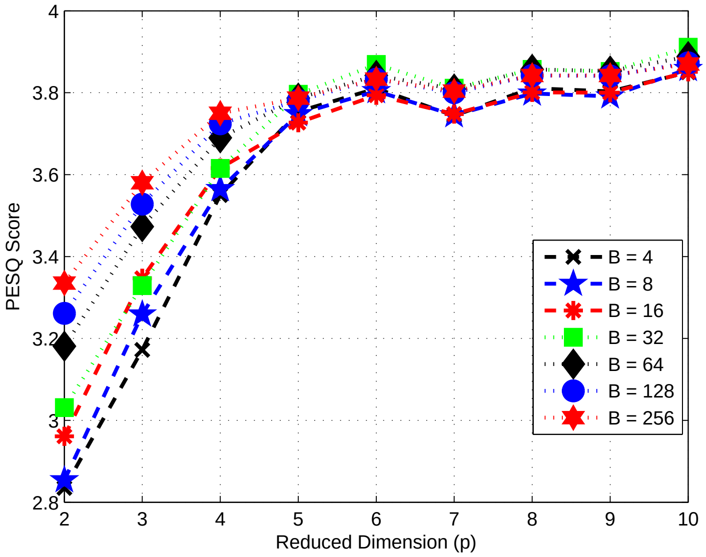


Figure 4.19: PESQ scores for SX set, $T = 0$, single index saved

Table 4.21: SegSNR scores for SA set, $T = 1$, single index saved

SegSNR(dB)										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	6.241	6.084	6.155	7.095	7.302	7.550	7.340	7.317	7.312	7.131
8	6.596	6.503	6.439	7.248	7.554	7.819	7.628	7.583	7.531	7.494
16	6.860	6.838	6.850	7.373	7.492	7.676	7.512	7.510	7.514	7.393
32	7.302	7.245	7.197	7.726	7.809	7.936	7.799	7.789	7.737	7.481
64	7.208	7.169	7.225	7.490	7.557	7.609	7.469	7.455	7.421	7.323
128	7.421	7.333	7.404	7.688	7.727	7.711	7.570	7.542	7.511	7.344
256	7.502	7.429	7.481	7.761	7.772	7.734	7.584	7.581	7.551	7.343

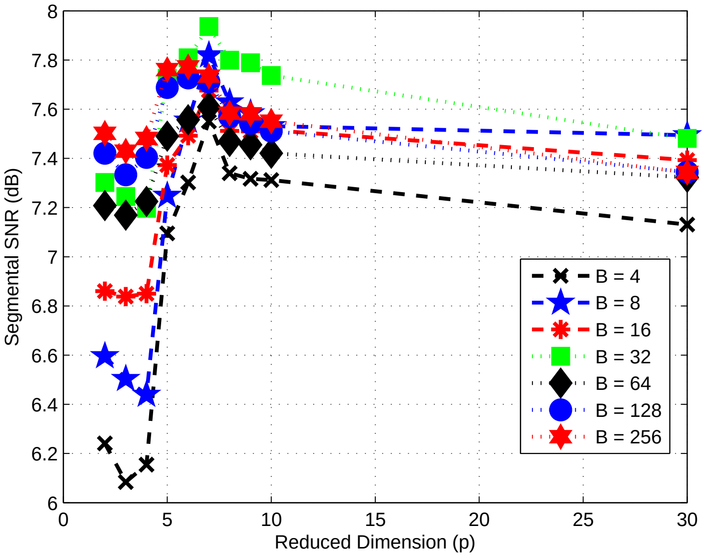


Figure 4.20: SegSNR scores for SA set, $T = 1$, single index saved

Table 4.22: SegSNR scores for SX set, $T = 1$, single index saved

SegSNR(dB)										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	6.862	7.116	7.264	8.125	8.067	8.159	8.679	8.768	8.311	8.847
8	6.945	7.108	7.160	7.820	7.783	7.839	8.320	8.037	7.708	8.281
16	7.324	7.365	7.423	7.949	7.938	7.937	8.240	8.022	7.687	8.247
32	7.468	7.417	7.613	8.263	8.249	8.269	8.615	8.534	8.148	8.683
64	7.989	7.909	7.805	8.189	8.124	8.137	8.355	8.271	7.998	8.314
128	8.141	8.090	8.093	8.405	8.351	8.355	8.562	8.625	8.232	8.525
256	8.331	8.280	8.201	8.472	8.415	8.415	8.536	8.540	8.184	8.280

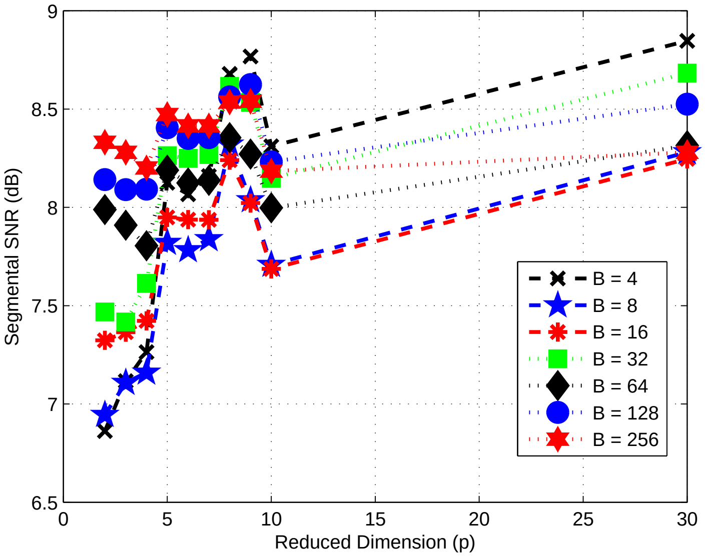


Figure 4.21: SegSNR scores for SX set, $T = 1$, single index saved

Table 4.23: LSD scores for SA set, $T = 1$, single index saved

LSD(dB)										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	5.189	5.025	4.956	4.753	4.512	4.413	4.373	4.337	4.289	3.859
8	4.833	4.683	4.664	4.508	4.278	4.180	4.154	4.123	4.074	3.652
16	4.520	4.401	4.381	4.246	4.067	3.985	3.963	3.933	3.903	3.515
32	4.394	4.283	4.267	4.140	4.001	3.935	3.899	3.874	3.850	3.468
64	4.177	4.102	4.084	3.971	3.863	3.803	3.774	3.757	3.735	3.377
128	4.049	3.984	3.971	3.867	3.769	3.714	3.687	3.671	3.652	3.317
256	3.987	3.931	3.918	3.822	3.737	3.682	3.656	3.642	3.626	3.305

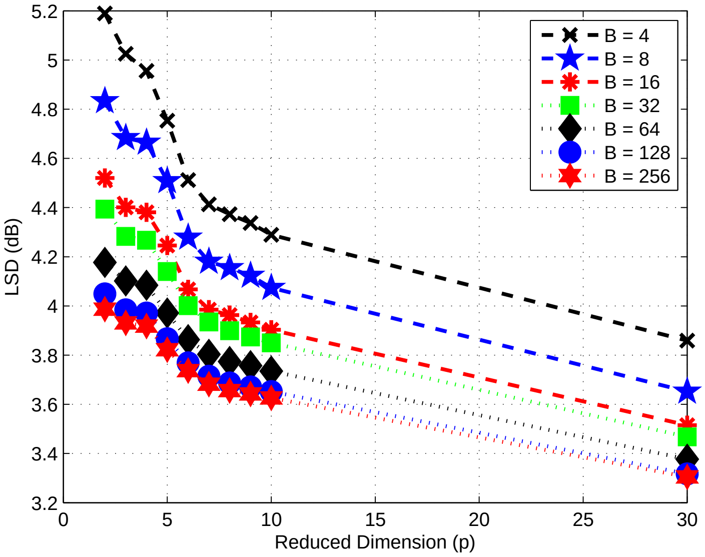


Figure 4.22: LSD scores for SA set, $T = 1$, single index saved

Table 4.24: LSD scores for SX set, $T = 1$, single index saved

LSD(dB)										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	5.405	5.279	5.065	4.907	4.865	4.824	4.684	4.409	4.292	3.843
8	5.146	5.022	4.829	4.667	4.619	4.556	4.439	4.182	4.059	3.643
16	4.739	4.647	4.493	4.370	4.336	4.282	4.200	4.001	3.890	3.504
32	4.650	4.563	4.416	4.305	4.278	4.227	4.150	3.959	3.906	3.464
64	4.387	4.317	4.203	4.111	4.088	4.045	3.982	3.824	3.772	3.365
128	4.195	4.139	4.034	3.955	3.936	3.903	3.851	3.771	3.662	3.286
256	4.079	4.027	3.937	3.865	3.848	3.832	3.780	3.697	3.592	3.244

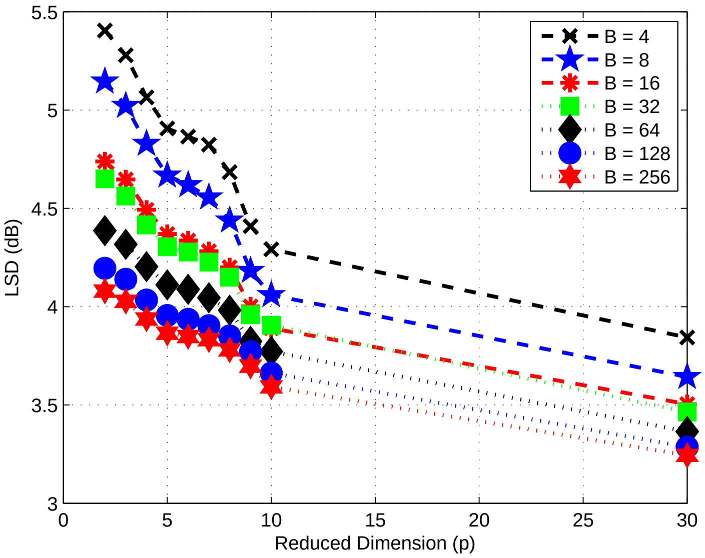


Figure 4.23: LSD scores for SX set, $T = 1$, single index saved

Table 4.25: PESQ scores for SA set, $T = 1$, single index saved

PESQ										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	2.444	3.586	2.603	3.112	3.486	3.590	3.603	3.604	3.602	3.875
8	2.939	3.914	2.941	3.192	3.535	3.673	3.674	3.675	3.679	3.952
16	3.119	3.894	3.095	3.325	3.622	3.697	3.704	3.713	3.715	3.950
32	3.175	3.853	3.153	3.356	3.618	3.695	3.703	3.710	3.713	3.966
64	3.304	3.795	3.287	3.458	3.631	3.684	3.692	3.697	3.700	3.908
128	3.364	3.849	3.347	3.499	3.663	3.720	3.729	3.730	3.735	3.928
256	3.412	3.855	3.400	3.540	3.677	3.728	3.656	3.736	3.737	3.902

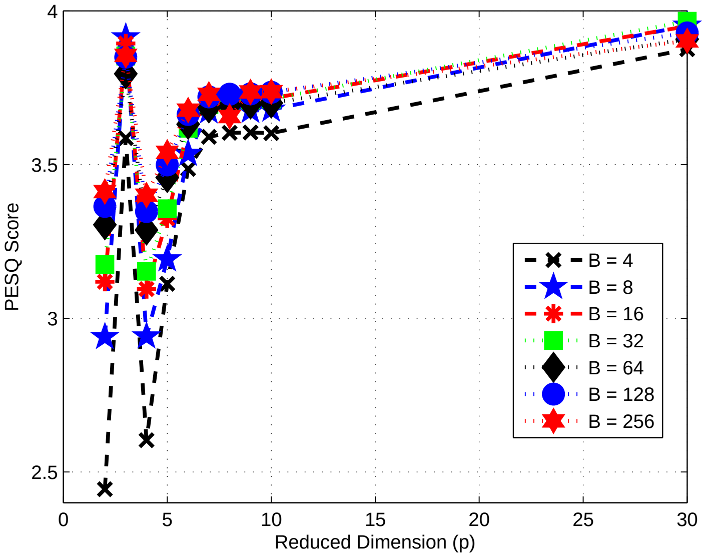


Figure 4.24: PESQ scores for SA set, $T = 1$, single index saved

Table 4.26: PESQ scores for SX set, $T = 1$, single index saved

PESQ										
B	Order of the estimation p									
	2	3	4	5	6	7	8	9	10	30
4	2.627	2.680	2.805	3.022	3.071	3.146	3.302	3.518	3.552	3.895
8	2.620	2.704	2.839	3.079	3.128	3.168	3.298	3.497	3.531	3.873
16	2.743	2.828	3.038	3.212	3.253	3.268	3.323	3.498	3.535	3.864
32	2.783	2.860	3.069	3.287	3.304	3.317	3.373	3.565	3.575	3.908
64	2.995	3.045	3.237	3.354	3.372	3.383	3.425	3.579	3.605	3.884
128	3.076	3.119	3.308	3.417	3.429	3.433	3.470	3.555	3.625	3.861
256	3.136	3.179	3.351	3.450	3.464	3.463	3.490	3.568	3.638	3.850

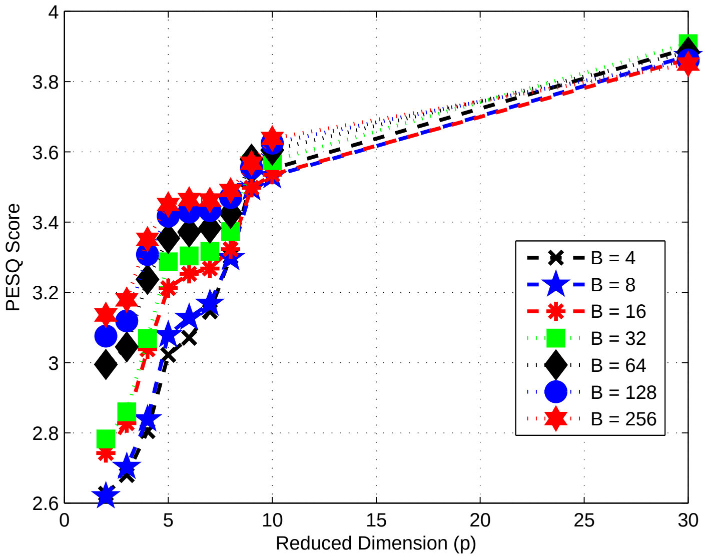


Figure 4.25: PESQ scores for SX set, $T = 1$, single index saved

4.5 Discussion of the Proposed System

We will first compare our model to the benchmark system, then we will look deeper into the performance effects of our model parameters individually.

4.5.1 Comparison with the Benchmark System

Both systems exploit the same excitaiton signal, therefore the comparison between speech files will address the comparison between the envelope extension methods. Benchmark system yields the best performance when the codebook size is 256, with an LSD of $\sim$ 4.6dB and a PESQ score of $\sim$ 3.3 over both SA and SX sets. Our model with the simplest parameters that give the worst performance with $B = 4$, $T = 0$ and $p = 2$ still yields a higher performance with an LSD of $\sim$ 4.2dB and a PESQ score of $\sim$ 3.7. It is possible to further improve the performance of our proposed model by changing variables such as increasing the codebook size or increasing the order of the estimation.

Similarly, our method outperforms similar HMM-based structures such as the one offered in [4]. The mentioned study obtains an LSD score of $\sim$ 7dB in extension band whereas our best results indicate a $\sim$ 3.2dB in overall wideband spectra.

4.5.2 The Effect of the Branch Number B

The temporal clustering in our system is conducted via HMMs and during the implementation, the branch number B of this HMM model is varied from 4 to 256 by the order of 2 in order to analyze the corresponding effect on the performance. Theoretically, increasing branch number B is expected to improve the performance of the model as it brings the freedom to choose from a larger set of codebooks, therefore making it easier to find a closer estimate to the original wideband envelope.

Indeed, the improvement with increasing branch number is visible especially with the LSD metric, which compare the original wideband envelope and estimated envelope. In almost every case, regardless of feature selection, dimension reduction, different data sets or temporal neighborhood presence; with increasing B we observe a monotonically decreasing LSD measure, which indicates a better estimation performance. For instance, Fig. 4.4 displays a 0.8 dB decrease is LSD for an estimation order p of 2 and a 0.6 dB decrease for an order of 10. This pattern is also apparent in Fig. 4.5. In Fig. 4.18, we observe the performance improvement on PESQ score as well. Therefore we can conclude that increasing the branch number B has a positive effect on our ABE model.

4.5.3 The Effect of Temporal Neighborhood

In order to observe the impact of temporal neighborhood on the performance of our method, we first compare two initial cases, where $T = 0$ and $T = 1$ with no feature selection introduced. Table 4.27 shows that introducing temporal information slightly increases performance over SA set.

However, we should note that Table 4.27 compares two different ordered estimators, order 10 and order 30, respectively. Indeed, if we compare two 10^{th} order filters where one exploits temporal neighborhood (therefore, feature selection as well) and the other does not; we see that introducing temporal neighborhood does not improve the performance, as can be seen by comparing Table 4.5 and Table 4.11. When $T = 0$, the LSD value varies between 3.918 and 3.341 but when $T = 1$ this interval rises to 4.025 and 3.358. Therefore, we can conclude that temporal neighborhood improves the performance of our system as long as the complete narrowband spectrum is used in the estimation. Otherwise, the information from neighboring frames does not bring any extra contribution to the performance.

Table 4.27: The effect of temporal neighborhood T

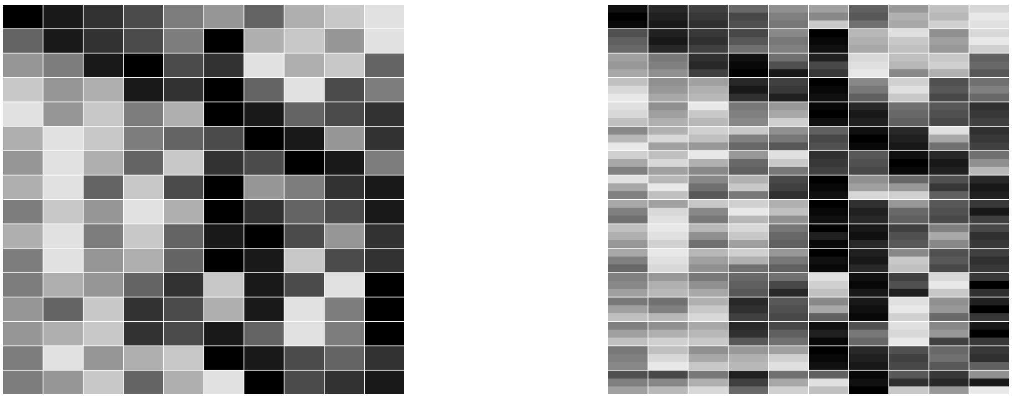
B	SegSNR (dB)		LSD (dB)		PESQ	
Branches	T=1	T=0	T=1	T=0	T=1	T=0
4	7.131	7.085	3.859	3.918	3.875	3.843
8	7.494	7.509	3.652	3.716	3.952	3.941
16	7.393	7.312	3.515	3.576	3.950	3.933
32	7.481	7.664	3.468	3.537	3.966	3.951
64	7.323	7.502	3.377	3.433	3.908	3.891
128	7.344	7.553	3.317	3.367	3.928	3.924
256	7.343	7.603	3.305	3.341	3.902	3.927

4.5.4 The Effect of Feature Selection

During linear estimation part of the proposed ABE system, instead of extracting $10(2T{+}1)^{th}$ order filters, we select the most correlated p instances from the R_{yx} matrix and decrease our order to p. The motivation of this selection process is the fact that mostly narrowband and wideband features are locally correlated.

In order to better explain this correlation structure, we present the cross-correlation matrix R_{yx} of a sample state in Fig. 4.26. In this visualization, darker regions indicate higher correlation and columns represent the narrowband LSF features whereas rows represent wideband LSF features. For $T = 1$, the temporal neighborhood is three frames and current time frame is represented by the center row. Analyzing this correlation matrix, we see that for lower frequencies, wideband LSF features and narrowband LSF features are linearly correlated. High frequency wideband features are also correlated with a set of high frequency narrowband features.

We observe that reducing the estimation order comes with the cost of a decrease in the estimation performance in terms of the LSD metric. In Fig. 4.22 and Fig. 4.23 the LSD increases by about $1 \sim 1.5$ dB as we decrease p from 30 to 2. Similarly, a smaller estimation order p degrades PESQ scores as well, as depicted in Fig. 4.25. SegSNR values do not follow a regular with varying dimension, though.



(a)$T = 0$ **(b)**$T = 1$

Figure 4.26: Sample R_{yx} correlation matrix visualizations. Columns represent the narrowband LSF features and rows represent wideband LSF features. The intensity of the gray level represents higher correlation.

In this context, the experimental results do not conflict with the theory; since by decreasing the estimation order we are weakening the precision of the estimation.

4.5.5 The Effect of Index Determinition

In the experimentation part, we implement and compare two cases with respect to the determinition of best correlated feature indexes where the first case model saves indexes for every state s_b while the second case model saves a single index list which is extracted from the average of R_{yx} matrices over all states. In theory, we expect the first case to yield better results due to its capacity to favor every state's character separately.

Experimental results do not conflict with the theory, and we observe a decrease in the performance when we decrease the number of index vectors in the memory, which is a fair trade-off.

4.6 Subjective Test Results and Evaluation

In addition to the objective tests, we have run a subjective A/B comparison test to assess the performance of our proposed model. The test requires the subjects to compare each A/B

Table 4.28: The subjective A/B pair comparison test results.

A	**B**	Score
Narrowband	B=256, T=0, p=10	1.12
Wideband	B=256, T=0, p=10	-1.40
256 CB size VQ	B=256, T=0, p=10	1.33
B=256, T=0, p=2	B=64, T=0, p=4	0.15
B=256, T=0, p=2	B=256, T=0, p=10	-0.01
B=256, T=0, p=2	B=256, T=1, p=30	0.09
B=64, T=0, p=4	B=256, T=0, p=10	0.00
B=64, T=0, p=4	B=256, T=1, p=30	-0.17
B=256, T=0, p=10	B=256, T=1, p=30	0.08
Identical Pair	Identical Pair	0.00

pair and express their preference on a scale of (-2; -1; 0; 1; 2) where the scale corresponds to *strongly prefer A, prefer A, no preference, prefer B* and *strongly prefer B*, respectively. The subjective A/B test includes 26 listeners, who compared 30 sentence pairs randomly chosen from our database. Of these 30, 3 pairs compared the proposed method with the narrowband version, 3 compared the proposed method with the benchmark system, 3 compared the proposed method with the original wideband version and 18 compared the proposed method with itself for varying p and T values. The final 3 pairs were identical. The results of this test are given in Table 4.28.

Subjective test results indicate that our model outperforms narrowband speech and the benchmark system significantly. In addition, the parameters of our model such as the branch number B or the size of the temporal neighborhood T do not reproduce audible differences.

Chapter 5

CONCLUSION

In this thesis, we focus on the Artificial Bandwidth Extension problem, which aims to map narrowband speech to wideband speech by estimating the missing upper band frequencies from the narrowband speech. Utilizing the well-known Source-Filter model, we propose a new system for the extension of the spectral envelope.

In Chapter 1, ABE problem is defined. This chapter also summarizes the previous work that is related to the subject and reveals the scope and structure of the thesis. Chapter 2 describes the Source-Filter model, with an emphasis on the feature representation of speech signals. Also included in this chapter are a review of the TIMIT database used during implementation, and a benchmark system using VQ. Chapter 3 contains our proposed system which introduces linear estimation into an HMM based model. In addition to a detailed description of the proposed system, brief introductions into Markov Models and Hidden Markov Models are also included. Chapter 4 presents the results of the implementation of the benchmark system and the proposed system in detail.

Our proposed model enriches an HMM based model with linear estimation. Our unsupervised multi-modal analysis framework starts with the multi-stream training of a parallel branch HMM model with a parallel corpus, which is composed of a wideband database and its narrowband counter part. Subsequently, given the multimodal feature vectors, we extract a state sequence associated with the feature vectors. When the state sequence is extracted, temporal clustering is complete and we continue with linear estimation. In linear estimation part, we define linear estimation filters from the narrowband source vector with temporal neighborhoods to estimate every single component of the wideband envelope, for every state. These estimation filters are calculated using Yule-Walker equations.

In the application part of the system, the multi-stream HMM model is decomposed into two models, representing narrowband model and wideband model, respectively, where both have same state transition probabilities. Therefore, for a given narrowband test frame, first, the state sequence is extracted with the narrowband model, and later, corresponding filters (with respect to the states) are used to map the narrowband envelope to wideband envelope.

We introduce several modifications to our model to investigate performance variations. The first modification is feature selection. Feature selection decides a finite number of most correlated instances within the narrowband source vector with the target wideband component. Once the instances are selected, a new source vector with these instances, which has a smaller dimension than the original source vector, is formed and similarly, corresponding estimation filters are estimated. Yet another modification to the model focuses on the instance selection process. With this modification, the most correlated instances are selected once over all frames instead of individually selecting instances for every other state.

In the experimentation, we observe the performance of our system thoroughly. First, three different objective metrics indicate that our method is significantly superior to the benchmark VQ system. We evaluate our system for varying parameters to examine the impact of parameters on the system and with the mentioned modifications, to decide whether they improve the performance of the system. Shortly, we have concluded that increasing the branch number in the parallel branch HMM model enhances the performance of our system, which is similar to VQ where increasing codebook size yields a better estimation. Incorporating temporal neighborhoods, on the other hand, brings a slight performance gain as long as all spectral components are used in the estimation. Feature selection reduces the order of the estimation at the expense of a degraded performance. During feature selection, utilizing a single index vector that is applicable to all states instead of calculating index vectors for every single state simplifies our model, therefore weakens the performance of the system.

In addition to the objective tests, we run an A/B comparison test to evaluate how the speech synthesized with our system is perceived and to see whether it is preferred to narrowband speech, original wideband speech or speech synthesized with benchmark system. The test contains comparison of the proposed system with itself for varying parameters as well. Subjective test results indicate that, even though not superior to original wideband speech, the wideband speech extended with our system is clearly preferred over the benchmark system and narrowband speech. The results also show that, changing the parameter does not produce audible differences at the synthesized speech signals.

5.1 Future Work

This study can be pursued further in following ways

i. The system uses LSFs as feature vectors. We may train and test the system using different representations of the spectral envelope.

ii. LSD measures (which compare envelopes only) matches the theory in almost every case whereas PESQ and SegSNR (which compare synthesized speech signals) occasionally fail to produce consistent results. This may be either due to the effect of extension of the excitation or due to the insufficiency of the PESQ and SegSNR distance metrics to evaluate the performance of the estimation. The former case may be avoided by employing a stronger excitation extension system and the latter case can be solved by evaluating the system with metrics more sufficient to the task.

Similarly, this framework of joint temporal analysis of correlated sources can be applied to any sufficient problem.

BIBLIOGRAPHY

[1] Peter Jax, "Enhancement of bandlimited speech signals: Algoriths and theoretical bounds," *PhD thesis, Institut für Nachrichtengeräte und Datenverarbeitung*, 2002.

[2] S. Voran, "Listener ratings of speech passbands," *IEEE Workshop on Speech Coding Proceedings*, pp. 81–82, September 1997.

[3] B. Iser and G. Schmidt, "Bandwidth extension of telephone speech," *Technical Report 2, Termic SDS Research*, June 2005.

[4] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, pp. 1707–1719, 2003.

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

[6] A. Buzo R. M. Gray Y. Linde, "An algorithm for vector quantizer design," *IEEE Trans. on Communications*, January 1980.

[7] K.Y. Park and H.S. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'00*, vol. 3, pp. 1843–1846, 2000.

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

[9] Yannis Agiomyrgiannakis and Yannis Stylianou, "Combined estimation/coding of high-band spectral envelopes for speech spectrum expansion," in *In: Proc. of INTERSPEECH*, 2004, vol. 1.

[10] A. Gersho, "Optimal nonlinear interpolative vector quantization," *IEEE Trans On Comm*, pp. 1285–1287, 1990.

[11] O. Cappe Y. Stylianou and Eric Moulines, "Continuous propabilistic transform for voice conversion," *IEEE Trans. Speech Audio Processing*, 1998.

[12] Laura Laaksonen Juho Kontio and Paavo Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, March 2007.

[13] L. R. Rabiner and R. W. Schafer, "Digital processing of speech signals," *Prentice Hall*, 1978.

[14] J. H. Hansen J. D. jr. and J. Proakis, "Discrete-time processing of speech signals," *Wiley-interscience*, 1993.

[15] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," *Prentice Hall*, 1993.

[16] A. Acero X. Huang and H. Hon, "Spoken language processing," *Prentice Hall*, 2001.

[17] Yu-Ren Luo Guido R.C. Shi-Huang Chen, "Speaker verification using line spectrum frequency, formant, and support vector machine," *11th IEEE International Symposium on Multimedia*, pp. 562–566, 2009.

[18] E. Erzin C. Yagli, "Artificial bandwidth extension of spectral envelope with temporal clustering," *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP. (Submitted)*, 2011.

[19] E. Erzin, "Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17/7, pp. 1316–1324, 2009.

[20] M. Sargin, E. Erzin, Y. Yemez, A. Tekalp, Erdem A., Erdem C., , and M. Ozkan, “Prosody-driven head-gesture animation,” *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'07*, vol. 2, pp. 677–680, 2007.

[21] L. F. Lamel W. M. Fisher J. G. Fiscus D. S. Pallett J. S. Garofolo and N. L. Dahlgren, “Darpa timit acoustic-phonetic continuous speech corpus,” *U.S. Department of commerce, nist speech disc 1-1.1 edition*, February 1993.

[22] ITU-T P.835, “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm,” *ITU-T Recommendation P.835*, 2003.

[23] Yi Hu and Philipos C. Loizou, “Evaluation of objective measures for speech enhancement,” in *In: Proc. of INTERSPEECH*, 2006.