INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

OBJECTIVE EVALUATION OF SPEECH QUALITY OVER TELECOMMUNICATION NETWORKS USING NEURAL NETWORKS

by

MOHAMED MOHAMED MEKY

A dissertation submitted to the Graduate Faculty in Engineering in partial fulfillment of the requirement for the degree of Doctor of Philosophy, The City University of New York.

1998

This manuscript has been read and accepted for the Graduate Faculty in Engineering in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

1/26/1998

Date:

Chairman of the Examining Committee:
Dr. Tarek N. Saadawi, Professor of Electrical Engineering, The City College of the City University of New York.

1/26/98

Date:

Executive Officer:
Professor Gerard Lowen, Dean of the Graduate Studies at the Engineering School, The City College of the University of New York.

Dr. Joseph Barbra — Professor, Department of Electrical Engineering, The City College of NY.

Dr. Mitra Basu — Associate Professor, Department of Electrical Engineering, The City College of NY.

Dr. Myung J. Lee — Assistant Professor, Department of Electrical Engineering, The City College of NY.

Dr. Thomas K. Mills — US Army Research Laboratory (ARL)

The City University of New York

Abstract

OBJECTIVE EVALUATION OF SPEECH QUALITY OVER TELECOMMUNICATION NETWORKS USING NEURAL NETWORKS

by

Mohamed Mohamed Meky

Advisor: Professor Tarek N. Saadawi

Throughout this research, we have introduced new perceptually objective measure techniques that can be used to predict speech quality over telecommunication systems. In these techniques, we emulate several known features of perceptual processing of speech sounds by human ear (including critical-band masking, equal loudness, and the intensity-loudness power law operations) to map the speech power spectrum into auditory power spectrum which assumed to represent the information conveyed by the auditory nerve to the brain. Then, we have used the neural network technique to map the effectively extracted features, that derived from the auditory spectrum, into the corresponding speech quality. The results indicate that our proposed techniques are reliable and robust in evaluating the coded speech quality and they are highly correlated to human responses across a wide range of quality levels and for a wide range of speech coding techniques.

From the speech coder designing point of view, our objective measure techniques can be used in development, testing, refinement, deployment, or standardization of algorithms and equipment that process speech signals.

Furthermore, in this research, we have presented a discussion of the issues involved in predicting the degradation impact of cell loss and jitter impairments on speech quality over ATM networks. Prediction the speech quality over ATM networks helps in designing

the speech coders and controlling their electrical parameters to maintain certain speech quality.

From the network design point of view, the proposed techniques can be used to assign upper cell loss and jitter limits for the suggested coding algorithms to be used over ATM networks. Also degradation information produced by the proposed techniques can be used to aid in designing of the management, congestion control protocols and assignment rules that achieve certain quality of service (QOS) requirements.

Dedication

My thesis is dedicated to the memory of my father Mohamed Moustafa Meky. Also, my thesis is dedicated to my mother, my parents in law, wife, sons, brothers, and sisters. And finally, I would like also to dedicate my thesis to my professors in Electrical Engineering Department, Faculty of Engineering, Alexandria University, Egypt, where I obtained my B. Sc. and M. Sc. degrees.

Acknowledgments

My thanks are wholly devoted to God who has helped me all the way to conclude this work successfully.

I am greatly indebted to my mentor Professor Tarek Saadawi for his gratitude, great help, continuous encouragement, and assistance during this research.

I would like to thank all the members of my Ph.D. committee: Professors Tarek Saadawi, Joseph Barbra, Mitra Basu, Myung J. Lee, and Dr. Thomas K. Mills for taking the time to review my thesis and making valuable comments to this work.

Special thanks to Dean Gerard Lowen for his support and the assistance I received from his office during my Ph.D. program.

I am very grateful to Dr. Peter Kroon, AT&T Bell Laboratories. He has supplied me with the speech data that allowed me to perform the computation work.

I wish to express my sincere thanks and gratitude to my wife for her kind encouragement during all stages of this research.

I would like also to express my deep and unlimited appreciation to my brother Moustafa Meky for his kind encouragement.

To all those and others who had helped me in one way or another, I express my gratitude.

Contents

Abstract iii

Dedication v

Acknowledgments vi

List of Figures xi

Chapter 1 INTRODUCTION 1

1.1 Overview 1

1.1.1 Subjective tests 1

1.1.2 Objective tests 2

1.2 Previous Work 3

1.3 Motivation 6

1.4 Work Performed 7

1.5 Structure of the Thesis 10

Chapter 2 A PERCEPTUALLY-BASED OBJECTIVE MEASURE FOR SPEECH CODERS USING ABDUCTIVE NETWORK . . 11

2.1 Overview 11

2.2 Speech Data 12

2.3 Calculation of the Bark Spectral Distance per Band (BSDB) 13

2.3.1 Preprocessing 13

2.3.2 Perceptual model 14

2.3.2.1 Critical band analysis 14

2.3.2.2 Equal-loudness preemphasis 16

2.3.2.3 Intensity-loudness power law 17

2.3.3 Bark Spectral Distance per Band (BSDB) 18

2.4 Abductive Network Concepts 20

2.4.1 Comparisons to neural networks 24

2.4.2 Comparisons to statistical regression 25

2.5 Evaluation System 25

2.6 Numerical Results 30

2.7 Validity of the evaluation system concept with conventionally-oriented parameters 35

2.7.1 Method 1 36

2.7.2 Method 2 37

2.8 Conclusion 39

Chapter 3 USING RADIAL BASIS FUNCTION NEURAL NETWORK IN PREDICTING THE SPEECH QUALITY 40

3.1 Overview 40

3.2 Calculation of the Perceptual Cepstrum Distance per Frame (PCDF) 41

3.2.1 Preprocessing 41

3.2.2 Perceptual Cepstrum Distance per Frame (PCDF) 42

3.3 Radial Basis Function Neural Network Concepts 43
3.3.1 Background 43
3.3.2 Network simulation 44
3.4 Evaluation System 45
3.4.1 Preparing the learning data set 45
3.4.2 Neural network's learning phase 47
3.4.3 Neural Network's test phase 48
3.5 Numerical Results 49
3.6 Conclusion 56
Chapter 4 DEGRADATION EFFECT OF CELL LOSS ON SPEECH QUALITY OVER ATM NETWORK 57
4.1 Overview 57
4.2 Impact of Cell Loss on Speech Quality 58
4.2.1 Numerical result 59
4.2.1.1 Replacement of the lost cell by silent samples 59
4.2.1.2 Replacement of the lost cell by the previous successfully received one 66
4.3 Conclusion 68
Chapter 5 IMPACT OF JITTER ON SPEECH QUALITY OVER ATM NETWORK 70
5.1 Overview 70
5.2 Calculation of the Perceptual LPC Parameters 72

5.3 Evaluation System 73
5.4 Validity of the Proposed Evaluation System 75
5.5 Impact of Jitter on Speech Quality 77
5.5.1 Numerical results 79
5.6 Conclusion 87
Chapter 6 CONCLUSION AND FURTHER WORK 89
Appendix A Linear Predictive Coding (LPC) Analysis 92
Appendix B Cepstral Analysis 94
REFERENCES 96

List of Figures

Figure 1.1.2.1	Methodology for development of objective quality assessment algorithm.	3
Figure 2.3.2.1	Block diagram of the perceptual model	14
Figure 2.3.2.2	Critical band masking curve	16
Figure 2.3.2.3	The weighting functions used for computing samples of the auditory spectrum B(z) from the power spectrum P(f).	18
Figure 2.3.3.4	Basic transformations used in obtaining BSDB(i).	19
Figure 2.4.5	Example five input, three layer abductive network.	20
Figure 2.4.6	The Predicted Squared Error.	23
Figure 2.5.7	Learning phase for one of the sub-evaluation system.	27
Figure 2.5.8	Selection one of the sub-evaluation system.	27
Figure 2.5.9	Predicted MOS error versus CPM for the evaluation system of Fig. 2.5.8.	28
Figure 2.5.10	The abductive network structure of the sub-evaluation system. The symbol b_i is the bark spectral distance per band i.	29
Figure 2.6.11	Actual and predicted MOS values for mixed speakers.	30
Figure 2.6.12	Actual and predicted MOS values for female speakers.	31
Figure 2.6.13	Actual and predicted MOS values for male speakers.	31
Figure 2.6.14	Actual and predicted MOS values for mixed speakers.	32
Figure 2.6.15	Actual and predicted MOS values for female speakers.	33
Figure 2.6.16	Actual and predicted MOS values for male speakers.	33

Figure 2.7.1.17 Critical band filters used in calculation the average signal-to-distortion ratio. 36

Figure 3.2.2.1 Basic transformations used in obtaining PCDF(l). 43

Figure 3.4.1.2 One of our speech files. S1 and S2 are denoted to the first and the second sentences respectively 45

Figure 3.4.1.3 Removing the silent period at the beginning of the two sentences: (a) sentence one, (b) sentence two. 46

Figure 3.4.2.4 Predicted square error for the learning phase of the radial basis functions neural network. 48

Figure 3.4.2.5 Learning phase for one of the sub-evaluation system. . . 48

Figure 3.5.6 Actual and predicted MOS values for sub-evaluation system 1. 49

Figure 3.5.7 Actual and predicted MOS values for sub-evaluation system 2. 50

Figure 3.5.8 Actual and predicted MOS values based on mixed speakers for the total evaluation system. 51

Figure 3.5.9 Actual and predicted MOS values based on female speakers for the total evaluation system. 51

Figure 3.5.10 Actual and predicted MOS values based on male speakers for the total evaluation system. 52

Figure 3.5.11 Actual and predicted MOS values for sub-evaluation system 1. 53

Figure 3.5.12 Actual and predicted MOS values for sub-evaluation system 2. 53

Figure 3.5.13 Actual and predicted MOS values based on mixed speakers for the total evaluation system. 54

Figure 3.5.14 Actual and predicted MOS values based on female speakers for the total evaluation system. 54

Figure 3.5.15 Actual and predicted MOS values based on male speakers for the total evaluation system. 55

Figure 4.2.1.1 Predicted MOS versus cell loss: Uniform distribution. . . . 60

Figure 4.2.1.2 Predicted MOS versus cell loss: Binomial distribution. . . 60

Figure 4.2.1.3 Predicted MOS versus cell loss: Poisson distribution. . . . 61

Figure 4.2.1.4 Degradation of speech quality versus cell loss: Uniform distribution. 62

Figure 4.2.1.5 Degradation of speech quality versus cell loss: Binomial distribution. 62

Figure 4.2.1.6 Degradation of speech quality versus cell loss: Poisson distribution 63

Figure 4.2.1.7 Degradation of speech quality of 128 kbps bit rate versus cell loss for different cell loss distributions 64

Figure 4.2.1.8 Degradation of speech quality of 32 kbps bit rate versus cell loss for different cell loss distributions. 64

Figure 4.2.1.9 Degradation of speech quality of 13 kbps bit rate versus cell loss for different cell loss distributions. 65

Figure 4.2.1.10 Average MOS values versus cell loss for different bit rate coding. 65

Figure 4.2.1.11 Average degradation in the speech quality versus cell loss for different bit rate coding. 66

Figure 4.2.1.12 Average MOS for a 32 kbps bit rate for the two replacement techniques. 67

Figure 4.2.1.13 Improvement effect of using the second replacement technique rather than using the first one. 68

Figure 5.2.1 Basic transformations used in obtaining the perecptual LPC parameters. 73

Figure 5.3.2 Predicted square error for the learning phase of the radial basis functions neural network. 74

Figure 5.3.3 Learning phase for one of the sub-evaluation system. . . 74

Figure 5.4.4 Actual and predicted MOS values for sub-evaluation system 1. 75

Figure 5.4.5 Actual and predicted MOS values for sub-evaluation system 2. 76

Figure 5.4.6 Actual and predicted MOS values for the total evaluation system. 76

Figure 5.5.7 Uniform jitter distribution. 78

Figure 5.5.8 Uniform jitter distribution. 78

Figure 5.5.9 Poisson jitter distribution. 79

Figure 5.5.1.10 Degradation of speech quality versus jitter: Uniform distribution. 80

Figure 5.5.1.11 Degradation of speech quality versus jitter: Binomial distribution. 80

Figure 5.5.1.12 Degradation of speech quality versus jitter: Poisson distribution. 81

Figure 5.5.1.13 Variation of speech quality versus jitter for bit rate = 32 kbps and 20% of cells face jitter. 82

Figure 5.5.1.14 Variation of speech quality versus jitter for bit rate = 32 kbps and 40% of cells face jitter. 82

Figure 5.5.1.15 Variation of speech quality versus jitter for bit rate = 128 kbps and 20% of cells face jitter. 83

Figure 5.5.1.16 Variation of speech quality versus jitter for bit rate = 128 kbps and 40% of cells face jitter. 83

Figure 5.5.1.17 Variation of speech quality versus jitter for 128 kbps bit rate. 84

Figure 5.5.1.18 Variation of speech quality versus jitter for 32 kbps bit rate. 85

Figure 5.5.1.19 Variation of speech quality versus jitter for 16 kbps bit rate. 85

Figure 5.5.1.20 Variation of speech quality versus jitter for 13 kbps bit rate. 86

Figure 5.5.1.21 Variation of speech quality versus jitter for 8 kbps bit rate. . 86

Figure A.1 Voice production model. 93

Chapter 1 INTRODUCTION

1.1 Overview

Great advances have been made in the telecommunications industry in the past decade. The next decade promises to be just as exciting with an explosion of new services and technologies. Now is the time to examine critical telecommunications issues. Specification and measurement of telecommunications system performance & its quality of service (QOS) is one of these critical issues [1]. Performance is important to end users when selecting telecommunications equipment/services and to providers when designing/operating telecommunications facilities. Among of these facilities are the speech coders. The primary motivation for characterizing the electrical performance of speech coders is that one hopes to maintain high speech quality by controlling their electrical parameters.

Subjective and objective tests are used to evaluate the performance of such coders.

1.1.1 Subjective tests

The purpose of the subjective listening tests is evaluate the speech coders performance. The most widely used of the subjective listening tests is the Absolute Category Rating (ACR) test. In an ACR test [2] the subjects listen to short stimuli and rate the quality on a five point scale. The rating scale used is shown in Table.1.1.1.1.

Description	Grade	Level of Distortion
Excellent	5	Imperceptible
Good	4	Just perceptible but not annoying
Fair	3	Perceptible and slightly annoying
Poor	2	Annoying but not objectionable
Bad	1	Very annoying and objectionable

Table 1.1.1.1 Quality rating scale for ACR test.

Listeners choose one of the descriptive labels shown in Table1.1.1.1. The selected response of each listener is transformed to the number shown and the arithmetic average over all listeners for a given test condition yields to Mean Opinion Score (MOS).

1.1.2 Objective tests

Objective measure techniques that provide reliable speech quality indications are indispensable to persons involved in the development, testing, refinement, deployment, or standardization of algorithms and equipment that process speech signals. Such techniques help to minimize the number of costly and time consuming formal subjective tests required by those activities [3], including the performance evaluation of the speech coders. To be generally applicable and useful, objective performance measures should be:

• Technology independent (independent of coding algorithm and transport architecture),

• Mimic the human perceptual response, so that objectively measured quality would agree with subjective quality.

There should be at least three major steps involved in the development of an algorithm for objective performance assessment. These steps are depicted graphically in Fig.1.1.2.1. First, subjective quality testing must be designed and conducted in such a manner that the results really do measure user-perceived quality. Second, candidate objective quality

measures are developed and implemented. The third major step is the validation of the objective quality assessment algorithm. A major criteria for the acceptance of candidate objective measures is accurate correlation to the user-perceived quality, as provided by the subjective quality ratings.

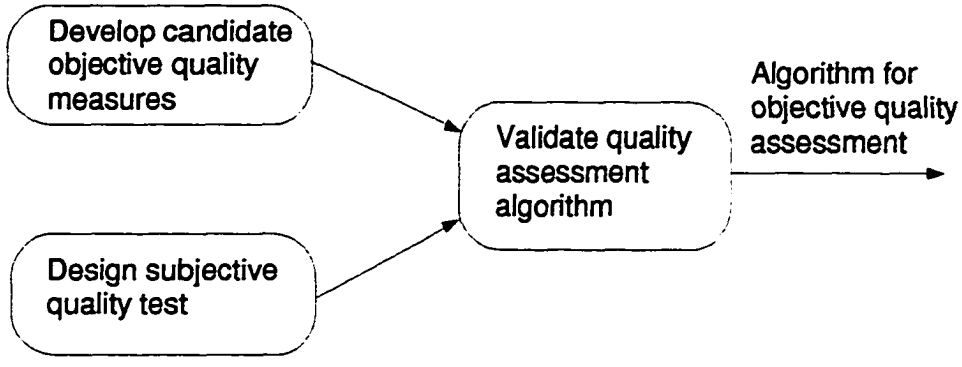


Figure 1.1.2.1 Methodology for development of objective quality assessment algorithm.

Objective measures that prove to be redundant or not significant may be discarded. The set of objective measures that accurately correlate with user-perceived quality levels can then be used in the design of the algorithm for objective quality assessment.

1.2 Previous Work

Over the years, there have been numerous objective measure suggested and used for the evaluation of speech coding systems. The most common objective measure is the mean squared error (MSE) between original and coded speech waveforms. This leads to signal-to-noise ratio (SNR), a term that is inaccurate because the "noise" does not exist as a physical entity; rather, it is manifested perceptually as the discrepancy between what the listener hears and what he expects to hear. The SNR treats the entire speech sample as a single vector, as if the listener made a single comparison after storing the

entire utterance, clearly an unreasonable proposition. In a more realistic variant called segmental (SNRSEG), SNR power ratios over short segments are determined and their geometric mean is computed. This seems to correspond much better to the auditory experience.

In an improved measure, a perceptually weighted mean squared error (PWMSE) was studied [4] where the weighting depends on the time-varing spectral envelope of the original speech. This measure is widely used for the analysis-by-synthesis in VXC/CELP coders. Recently, it has also been proposed as the basis for a modified SNR measure [5]. However, even with perceptual weighting such a distortion measure is still fundamentally waveform-based and focuses on approximating the sample-to-sample variations of the signal, rather than on the perceptual closeness of the reconstructed speech to the original [6].

Other objective measures are derived from LPC coefficients or other time/frequency domain parameters [7, 8, 8] but none of these have been totally satisfactory. In particular, the cepstral distance (CD) measures the disparity between the original and coded spectral envelopes, and is determined by generating cepstral coefficients from the LPC parameters. In [7], a high correlation coefficient of 0.93 was claimed between CD and MOS for speech coded in the range of 16-32 kb/s. However, for a broader range of bit rates, Quackenbus et al. [9] reported a much lower correlation of 0.63.

The limited ability of some of these measures to predict MOS ratings reported in [9] is shown in Table 1.2.2, where r is the correlation coefficient between MOS ratings and corresponding objective measures. The value r = 1 would indicate that the measure predicts MOS perfectly, while r = 0 could be obtained even by randomly guessing

the MOS. As is seen, conventional SNR does not perform much better than this, but others do. For example, segmental SNR reaches 0.77, with LPC-based measures trailing closely behind it.

Objective Speech Quality Measure	r
Time domain measure	
SNR	0.24
Segmental SNR	0.77
LPC-based measures	
Log area ratio	0.62
Log likelihood ratio	0.50
Cepstral distance	0.63
Frequency variant log spectral distance	
LPC-based	0.68
Filter bank	0.72
Weighted-slope spectral distance	0.74

Table 1.2.2 Comparative performance of objective measures.

Objective measures based on auditory models [10, 11] are also used in the field of speech recognition. Yet their goal there is different: to measure the phonetic distance between received and stored utterances, usually without regard to the speakers's identity. In contrast, high-quality speech coding demands the minimization of psycho-acoustic distance, of which speaker identity is an integral part.

The performance of some objective methods is quite promising such as the bark spectral distance (BSD) [6], and information index [12]. However, a technique accurate and robust enough to replace human listeners in all situations is not currently feasible [1]

1.3 Motivation

While the quality of classical waveform coding algorithms can be estimated by waveform matching via the signal-to-noise ratio (SNR) or the segmental SNR, these measures are of little relevance to the new generation of coding algorithms in the range of 2-8 kb/s where exact waveform preservation would be an unrealistic goal. Instead, such coders aim for the adequate rendering of only the perceptually significant aspects of the signal to preserve intelligibility and naturalness. Consequently, for these coders we must often depend on informal listening (subjective test) to make instant judgments of quality during the process of algorithm development. This is, unfortunately, a hazardous procedure since casual listening does not reliably reveal whether one version of an algorithm is better or worse than a slightly different alternative. Also, once a candidate algorithm has been developed there is no simple and reliable way to report on its overall quality short of the costly and time-consuming process of formal subjective testing.

The degradation introduced by different types of low-rate coders and by the effects of transmission errors on such coders are very diverse. It is indeed a challenging task for any one listener to predict how two coders with different kinds of distortion will rank in formal subjective tests.

Finding good objective quality measures for coded, distorted, or synthetic speech is one of the most important problems in speech processing [9]. If an objective quality measure could be found which was highly correlated with the results of human preference tests, its utility would be undeniable. For example, it could replace subjective quality measures in the design of speech coding, synthesis and communication systems, in which the measure could be repeatedly applied in the course of the design so as to iteratively optimize the

system's parameters. Alternatively, it could be an integral part of a speech system itself, providing the fidelity criteria to be dynamically optimized as part of the coding procedure. Relative to subjective tests, objective tests are less expensive to administer, give more consistent results, and are not subject to the human failings of administrator or subject. And since objective test results are consistent, tests applied at different times and with different personnel and testing facilities could be directly compared. This is not generally the case for subjective test results. Of course, objective tests cannot eliminate the need for subjective testing but it can greatly aid the algorithm development process by allowing a preliminary evaluation of alternative candidate coders. Also they are indispensable to end users when selecting telecommunication services.

Furthermore, the reliable objective measure techniques can be used to predict the degradation impact of cell loss and jitter impairments on speech quality over ATM networks. Prediction the speech quality over ATM networks helps in designing the speech coders and controlling their electrical parameters to maintain certain speech quality.

From network design of view, the reliable objective techniques can be used to assign upper cell loss and jitter limits for the suggested coding algorithms to be used over ATM networks. Also degradation information produced by these techniques can be used to aid in designing of the management, congestion control protocols and assignment rules that allowed the meeting of connection performance standards and the achieving of certain quality of service (QOS) requirements.

1.4 Work Performed

In most previous objective measures, speech quality is estimated by measuring the distortion between input and output speech records, and using regression to map the

distortion values into estimated quality.

In this research, we introduce our objective measure techniques that emulate several known features of perceptual processing of speech sounds by human ear (including critical-band masking, equal loudness, and the intensity-loudness power law operations) to map the speech power spectrum into the corresponding auditory power spectrum (bark domain). Then, the auditory power spectrum has been processed by the following speech operations that summarized, for the first objective technique, as follows:

• Calculate the bark spectral distance per band (BSDB) between the input and the output coded speech signals,

• Use the abductive networks, that evolved from neural network, statistical modeling, and artificial intelligence concepts, to estimate the speech quality from the BSDB.

While the speech operations in the second objective algorithm are:

• Derive the perceptual LPC coefficients from the auditory spectrum that used to calculate, for each frame, the cepstrum distance between the input and the output coded speech signals,

• Use the radial basis functions neural network to map the perceptual cepstrum distance per frame into the corresponding estimated speech quality.

After extensive experimentation and validation of our techniques, we obtained results of r and s in the range of 0.96 and 0.1 respectively where r is the correlation coefficient between subjective test ratings and corresponding objective measures, and s is the standard deviation of the prediction error. The results indicate that our proposed techniques are reliable and robust in evaluating the coded speech quality and they are highly correlated to human responses across a wide range of quality levels and for a wide range of speech

coding techniques.

Furthermore, in this research, we have presented a discussion of the issues involved in predicting the user's opinion of speech quality due to cell loss and jitter introduced in ATM networks. The user's opinion is important for the proper design of network algorithms such as routing, flow control, and management techniques.

We have used the first objective technique in studying the impact of cell loss on speech quality over ATM networks. Moreover, we compare the results between two different cell loss's replacement techniques: stuffing silent samples and inserting the previous information in the lost cell. Study shows that the second replacement techniques produces better result when compared with the first one. The study also shows that up to 10% of speech cells can be lost over ATM networks while keeping the speech quality over MOS (Mean Opinion Score) of 3.2 for some speech coders.

Since introducing jitter for the speech signals will change their time characteristics, we can't use the previous objective techniques in predicting the ATM network's output speech quality because the time matching between the input and the output signals is necessary in that techniques. To study the impact of jitter on speech quality, we have introduced a new perceptually-output-based objective technique to predict the output speech quality without referring to the input speech. In this technique, we have used the auditory power spectrum in deriving the perceptual LPC parameters. Then, we use the radial basis functions neural network to map the perceptual LPC into the corresponding estimated speech quality.

From speech coder designing point of view, prediction the speech quality over ATM networks helps in designing the speech coders and controlling their electrical parameters

to maintain certain speech quality.

From network design of view, the proposed techniques can be used to assign upper cell loss and jitter limits for the suggested coding algorithms to be used over ATM networks. Also degradation information produced by the proposed techniques can be used to aid in designing of the management, congestion control protocols and assignment rules that allowed the meeting of connection performance standards and the achieving of certain quality of service (QOS) requirements.

1.5 Structure of the Thesis

In the following chapter, chapter 2, the first objective measure technique is presented. In chapter 3, we introduced the second objective measure technique. Studying the degradation impact of cell loss on speech quality over ATM networks is introduced in chapter 4. In the fifth chapter, we have presented the issues involved in predicting the degradation impact of jitter on speech quality over ATM networks and finally, chapter 6 contains our concluding remarks.

Chapter 2 A PERCEPTUALLY-BASED OBJECTIVE MEASURE FOR SPEECH CODERS USING ABDUCTIVE NETWORK

2.1 Overview

The purpose of this chapter is to provide a new perceptually-oriented objective measure technique that is highly correlated to human responses across a wide range of quality levels and for a wide range of speech processing, transmission, and transport technologies [13]. In this chapter, we introduced our objective measure technique that:

• Emulates several known features of perceptual processing of speech sounds by human ear (including critical-band masking, equal loudness, and the intensity-loudness power law operations) to map the speech power spectrum into auditory power spectrum (bark domain) that is used to calculate the bark spectral distance per band (BSDB) between the input and the output coded speech signals.

• Uses the abductive networks, that evolved from neural network, statistical modeling, and artificial intelligence concepts, to estimate the speech quality from the BSDB.

After extensive experimentation and validation of our technique, the results indicate that our proposed technique is reliable and robust in evaluating the coded speech quality. A major achievement of our objective measure technique is the realization of a speech evaluation method that is technology independent (independent of coding algorithm and communication technology).

The rest of this chapter is organized as follows: section 2 introduces our speech data. The calculation details of the bark spectral distance per band (BSDB) are explained in section 3. Fourth section shows the abductive network concepts. The proposed evaluation

system is introduced in section 5. Section 6 summarizes the numerical results. Validity of the evaluation system concept with conventionally-orinted parameters is presented in section 7 while section 8 contains the conclusions.

2.2 Speech Data

Testing and evaluation of the proposed approach is based on speech files database [14] which has two types of speech material, unfiltered (flat) and IRS filtered speech. The speech material consists of sentence pairs by 3 male and 3 female speakers (4 pairs per speaker). A total of 44 listeners participated in the test. Each group of 11 listeners would listen to all talkers, but only one of the 4 sentence pairs. The listening was done over headphones using playback over one ear. The listeners have normal hearing and non are experts in speech processing technology and they are given a small number of practice trials before the experiment starts.

For our objective measure technique, we use the source speech files and speech files (24 * 24 files) that processed by 24 coders in development and testing our evaluation system. These coders cover a wide range of quality levels and a wide range of speech processing and transmission technologies. The results of the listening tests for a variety of bit rate coders, that used the IRS filtered speech, are listed in Table 2.2.1.

Coder #	Bit-rate Kb/s	Remark	MOS
1	32		3.77
2	32	With silence coding	3.82
3	16	Floating point	3.88
4	16	10 % fe. rate	2.27
5	16	10 % fe. rate	2.31
6	16	10 % fe. rate	2.79
7	16	3 % fe. rate	2.26
8	16	3 % fe. rate	3.58
9	16		3.73
10	16	With silence codi	3.9
11	16		3.79
12	16		3.82
13	13		3.63
14	8		3.56
15	8		3.33
16	7.95		3.49
17	6.8	Floating point	3.4
18	6.8	Fixed point	3.47
19	6.25/5.8		3.2
20	5.6		3.33
21	4.8		3.03
22	2.4		2.7
23	2.4		2.73
24	1 - 8	Float, variable rate	3.45

Table 2.2.1 Results of the subjective test.

2.3 Calculation of the Bark Spectral Distance per Band (BSDB)

2.3.1 Preprocessing

As a first step, the input and output records are aligned in time. The relative delays are determined by cross correlating the input and output speech envelops. To avoid system gain effects, the records are corrected to have equivalent average power in the speech periods. The speech frame of 10 ms is weighted by Hamming window and the consecutive frames overlapped by 50%. If in a given frame the signal was found to fall below a threshold power level, the contribution of that frame (silent period) to the

average distortion was set to zero. By computing the magnitude square FFT spectrum, the frame's power spectrum P(f) is calculated and followed by several stages.

2.3.2 Perceptual model

In this technique, we would like to emulate several known features of perceptual processing of speech sounds by the human ear to map the speech power spectrum, P(f), into auditory power spectrum, B(z), which is assumed to represent the information conveyed by the auditory nerve. A block diagram of the perceptual model is shown in Fig. 2.3.2.1.

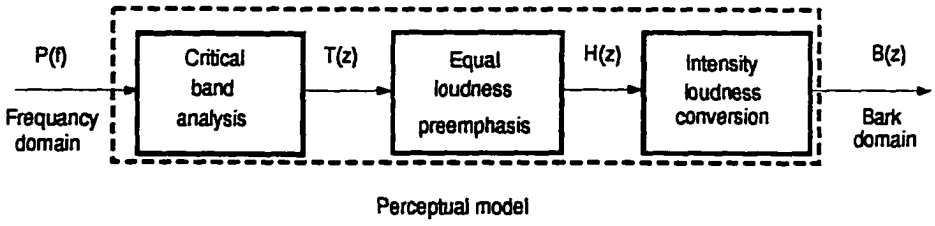


Figure 2.3.2.1 Block diagram of the perceptual model

2.3.2.1 Critical band analysis

The most important part of the ear is the tiny snail-shaped bony structure in the inner ear called the cochlea. Inside the cochlea is a highly sophisticated mechanism for converting the incoming pressure variations into the electrical signal of the auditory nerve [15]. Of particular interest is the function of the basilar membrane and the corresponding transduction process of the hair cells in the inner ear [16], [17]. This process causes hearing perception to be a function of the frequency of the sound, with frequency sensitivity depending on the position along the basilar membrane (the place theory) [18]. It is known that within some critical region of frequency, sound energies are interacting with one another. As we move outside the critical region, sound energies no longer

interact. The critical region is called the critical band [15]. The procedure of converting Hz to Bark follows the established view of auditory perception in psychoacoustics, which holds that the frequency-to-place transformation along the basilar membrane of the inner ear is in terms of critical bands whose bandwidth is one Bark [19]. Thus as a first step, the power spectrum P(f) is warped along its frequency axis, f, into the bark frequency, z, to obtain what is called "critical-band density" [4], P(z), via the relation [20]:

$$f = 600 \ sinh\,(z/6) \tag{2.3.2.1}$$

where f is the frequency in Hz. As hearing perception is directly related to the deformation of the basilar membrane, caused by different frequency components, it is not surprising that sound components at particular frequencies may reduce the perceptual effects of other sound components at neighboring frequencies. This effect is called masking [8]. Masking is thought to result from the spread of the input energy along the length of the basilar membrane [21]. The invariance of the shape of the masking characteristics with changes in masker frequency when plotted on a critical band rate scale (in Bark) [22] gives us another advantage of using the bark scale. The approximate power spectrum of the simulated critical-band masking curve $\Psi(z)$, shown in Fig. 2.3.2.2, is given by [23]:

$$\Psi(z) = \begin{cases} 0 & for\ z < -1.3\ , \\ 10^{2.5(z+0.5)} & for\ -1.3 \le z \le -0.5\ , \\ 1 & for -0.5 < z < 0.5\ , \\ 10^{-(z-0.5)} & for\ 0.5 \le z \le 2.5, \\ 0 & for\ z > 2.5 \end{cases} \tag{2.3.2.2}$$

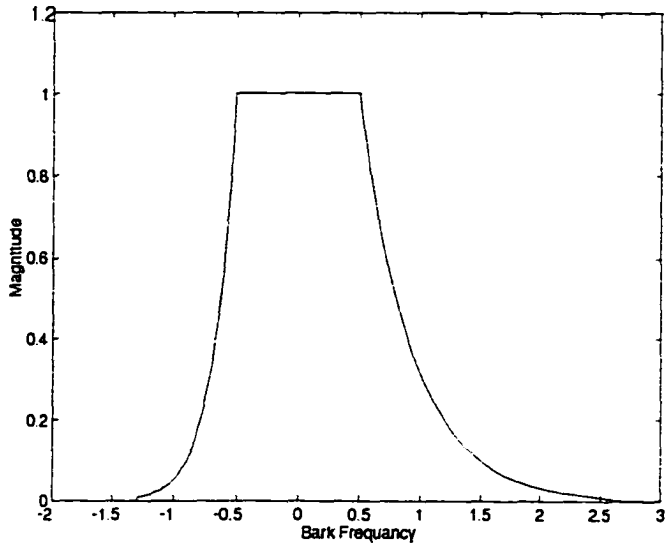


Figure 2.3.2.2 Critical band masking curve

The application of the auditory filter to derive the neural excitation pattern, T(z), is accomplished by convolving the critical-band density, P(z), with the critical-band masking curve, $\Psi(z)$. The excitation pattern primarily models the auditory nerve response to vowel sounds [19]. Smoothing speech spectrum of high dimensionality, such as yielded by FFT, by the critical-band masking curve $\Psi(z)$ leads to a spectral of lower dimensionality. Thus the excitation pattern, T(z), is sampled in approximately 1-bark intervals. Typically, 17 spectral samples of T(z) are used to cover the 0–15.575 bark (0–4 kHz) analysis bandwidth (with sample 1 equal to sample 2 and sample 16 equal to sample 17 [23]).

2.3.2.2 Equal-loudness preemphasis

In this stage, the threshold of hearing and the nonlinear frequency response of the ear to intensity differences are taken into account. This effect is calculated by multiplying the samples of the excitation pattern T[z(f)] by the simulated equal-loudness curve, E(f)

[23]:

$$H(z(f)) \;=\; E(f)\,T[z(f)] \tag{2.3.2.3}$$

The function E(f) is an approximation to the non-equal sensitivity of human hearing at different frequencies [24]and simulates the sensitivity of hearing at about the 40-dB level. According to [23], E(f) is given by:

$$E(f) \;=\; \frac{\left[(2\pi f)^2 \;+\; 56.8 * 10^6\right](2\pi f)^4}{\left[(2\pi f)^2 \;+\; 6.3 * 10^6\right]^2 \left[(2\pi f)^2 \;+ 0.38 * 10^9\right]} \tag{2.3.2.4}$$

2.3.2.3 Intensity-loudness power law

As a last stage, the samples of the auditory power spectrum B(z) is given by applying cubic-root amplitude compression of H(z) [22]:

$$B(z) = [H(z)]^{0.33} \tag{2.3.2.5}$$

This operation simulates the nonlinear relation between sound intensity and perceived loudness.

In summary, the perceptual model takes into account the human ear's nonlinear transformations of frequency and amplitude, together with important aspects of its frequency analysis and masking behavior in response to complex steady-state sounds.

Following Hermansky's method [23], the convolution of P(z) with critical band $\Psi(z)$ and the effect of the equal loudness curve are carried out for each sample of H(z) in the frequency domain with a precomputed weighting functions, covering the (0–4 kHz) spectrum, shown in Fig. 2.3.2.3. Referring to eq. 2.3.2.5, we get the samples of auditory spectrum , B(i).

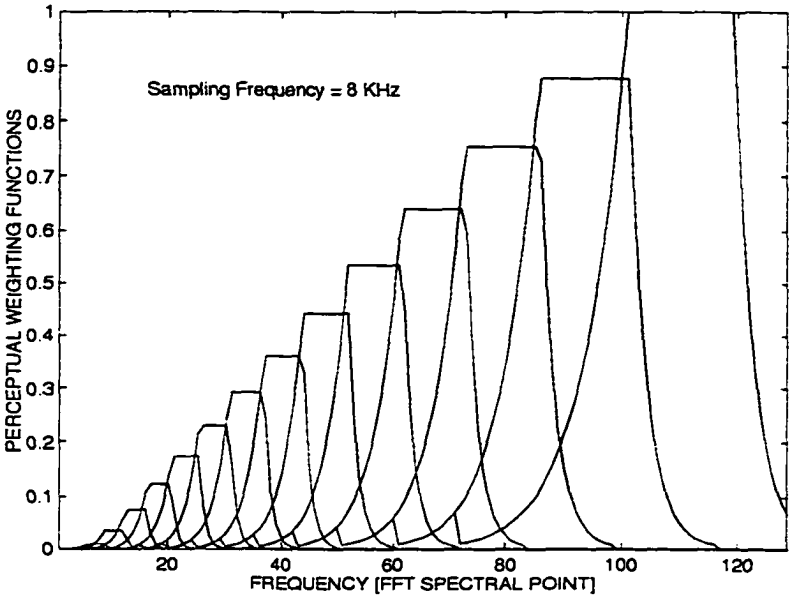


Figure 2.3.2.3 The weighting functions used for computing samples of the auditory spectrum B(z) from the power spectrum P(f).

For telephony application, we used thirteen samples of the auditory spectrum, B(z), to cover the spectrum from 300–3400 Hz.

2.3.3 Bark Spectral Distance per Band (BSDB)

The auditory spectrum B(z) reflects the ear's nonlinear transformations of frequency and amplitude, together with aspects of its frequency analysis and spectral integration properties in response to complex sounds [6]. For each band, i, the square Euclidean distance between the auditory spectrum of the input and the output is given by:

$$dis\left\{B_x^l(i), B_y^l(i)\right\} = \left[B_x^l(i) - B_y^l(i)\right]^2 \qquad (2.3.3.6)$$

where $B_x^l(i)\,and\,B_y^l(i)$ are the auditory spectrum samples of frame l of the input and output speech respectively. The bark spectral distance per band, BSDB (i) is given by:

$$BSDB(i) = \frac{\sum_{l=1}^{N} dis\{B_x^l(i)\,.\,B_y^l(i)\}}{\sum_{l=1}^{N}\sum_{i=1}^{b}\left[B_x^l(i)\right]^2} \qquad (2.3.3.7)$$

where N is the number of frames in the utterance while b is the number of bands. Figure 2.3.3.4 illustrates the basic transformations used in obtaining BSDB (i).

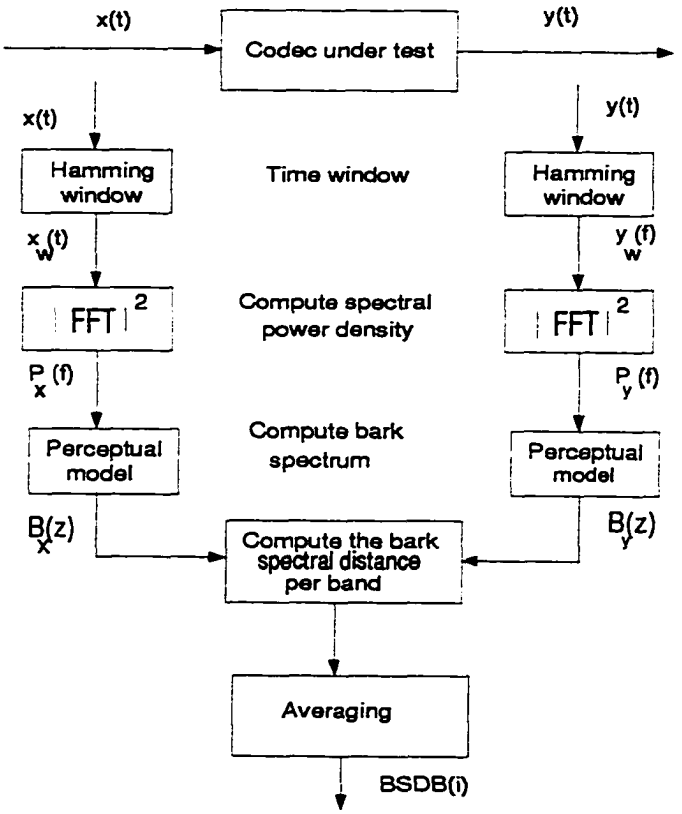


Figure 2.3.3.4 Basic transformations used in obtaining BSDB(i).

2.4 Abductive Network Concepts

Abductive reasoning is defined as the act or process of reasoning from principles to facts under uncertainty using numeric measure and functions [25]. Because of the inherent uncertainty associated with abduction, analytic solutions are generally precluded and empirical methods such as induction must be employed. Abductory induction is a special class of inductive reasoning for synthesizing abductive principles (functions) from database of empirical observations [25]. An important class of abductive functions are abductive networks. An abductive network is a set of interconnected function nodes. information flows from the original input variables through the network to the output variables. Figure 2.4.5 shows an example of a five input, three layer abductive network.

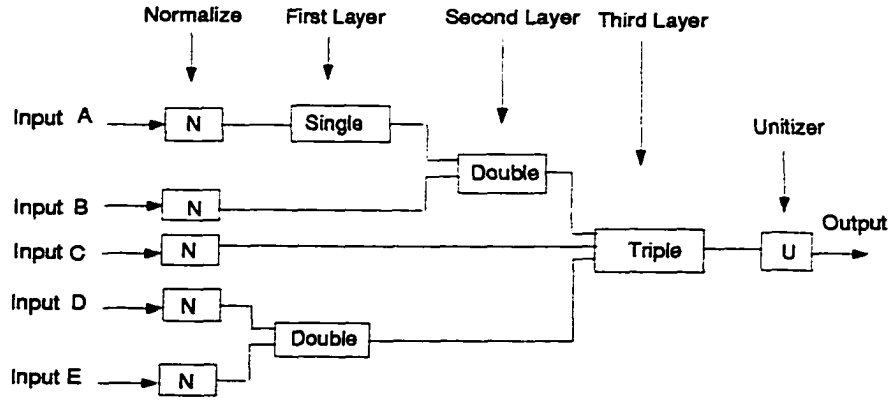


Figure 2.4.5 Example five input, three layer abductive network.

The functional element coefficients, number of network elements, types of network elements, and the connectivity are learned from the data. There are seven type of nodes or elements [26].

The algebraic form of each element is shown in the equations below where w_n are the weighting coefficients and x_n are the input variables.

1. Singles: $w_0 + w_1x_1 + w_2x_1^2 + w_3x_1^3$

2. Doubles: $w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 + w_6x_1^3 + w_7x_2^3$

3. Triples: $w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_1^2 + w_5x_2^2 + w_6x_3^2 + w_7x_1x_2 + w_8x_1x_3 + w_9x_2x_3 + w_{10}x_1x_2x_3 + w_{11}x_1^3 + w_{12}x_2^3 + w_{13}x_3^3$

Singles, doubles, and triples are elements whose names are based on the number of input variables.

4. Normalizers: $w_0 + w_1x_1$

Normalizers transform all of the original input variables into a relatively common region with a mean of zero and a variance of one.

5. Unitizers: $w_0 + w_1x_1$

A unitizer converts the range of the network outputs to a range with the mean and variance of the output values used to train the network.

6. White elements: $w_1x_1 + w_2x_2 + ... + w_nx_n$

The white element consists of the linear weighted sums of all the outputs of the previous layer.

7. Wire elements: The wire element is used for a network that consists only of a normalizer and a unitizer.

Research based on the paradigm of neural networks has led to self-organizing and adaptive methods in modeling and classification algorithms [27]. One very effective algorithm for inductively creating networks is called the Abductory induction mechanism (AIM) [25]. The AIM has been developed drawing from almost three decades of neural network

and algebraic modeling research by many authors [28, 29]. it is a supervised inductive learning tool that uses non-parametric regression to generate complex decision, prediction, and control models. The algorithm makes few a-priori assumptions about the structure of the process it is modeling. it first attempts to create standard 1-3 variable polynomial equations (called elements) using combinations of the original input variables. Then it looks for ways to combine the resulting nodes with each other and/or other original input variables to create new, higher order nodes. This network synthesis process continues, building the network layer by layer, until no additional process can be made. Of all the nodes that were created during synthesis, the node which represents the "best" model is selected as the final output node. All nodes that are not along the path from the output node to the original input variables are discarded, leaving the final connected network. Throughout the synthesis process, the algorithm makes value judgements concerning which nodes are well suited for modeling and which are less attractive by using an advanced statistical modeling criterion. The modeling criterion used within the abductive network is the predicted squared error (PSE) criterion derived by A. R. Barron [30]. The PSE is a heuristic measure of the expected network squared error for independent data not in the training database. In AIM, the PSE is given:

$$PSE = FSE + KP \tag{2.4.8}$$

where FSE is the fitting squared error of the model on the training data and KP is a complexity penalty. Figure 2.4.6 shows the relationship between the FSE, PSE, and KP, the Complexity Penalty, which given by [30, 31]:

$$KP = CPM * \frac{2K}{N_1} s_p^2 \tag{2.4.9}$$

where K, N_1, and s_p^2 are determined by the database of examples used to synthesize the network and K is the total number of coefficients, N_1 is the number of training data, s_p^2 is an a-priori estimate of the true unknown model error variance, and CPM, the Complexity Penalty Multiplier, is a chosen numerical value. Experience has verified that the choice of $s_p^2 = \sigma_0^2 / 2$ gives acceptable results [30] where σ_0^2 is the variance of the output values used in training the network. A higher value for the CPM increases the impact of the complexity penalty term, which will result in a less complex network, while a lower value for the CPM decreases the complexity penalty impact and results in a more complex network.

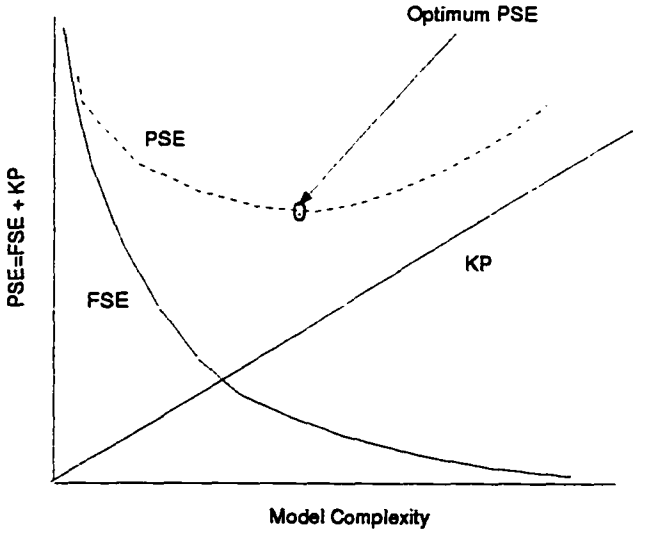


Figure 2.4.6 The Predicted Squared Error.

The algorithm introduces the proper network structure and coefficients by minimizing PSE criterion that attempts to select as accurate a network model as possible without overfitting the data. Overfitting occurs when the network model becomes so specific to

the training data that it does a poor job of modeling new data. The technique minimizes overfit by performing a trade-off between model complexity and accuracy, based on the assumption that simpler models are more general and superior for as yet unseen data (i.e. data not used for training) [19].

2.4.1 Comparisons to neural networks

Neural networks are a powerful technology, based on analogies to biological neurons, for learning relationships from databases of examples. The ability of abductive networks to model and solve complex problems is not because of the massive parallelism or the similarity of network nodes to neurons. The power of abductive networks is derived from the fundamental concepts that functions are a powerful knowledge representation and that networks subdivide complex problems into many simpler ones, greatly simplifying machine learning. These facts result in a practical means for computers to deal with uncertainty, learn from examples, and handle extremely complex relationships in a compact and rapidly executable form [25]. This networking concept is directly related to the concept of "chunking" proposed by George Miller [32]. Miller observed that humans are only able to handle seven, plus or minus two, item effectively at any one time. As Miller explains, the way people get around this processing constraint is to chunk, or group together, a number of elements and treat the group as a unit. Although abductive networks evolved from neural network research , it has become substantially different than neural network. A key difference is that abductive networks have fewer, more powerful nodes than neural networks. Abductive network also differs from neural network tools because it uses advanced statistical methods and applies a modeling criterion to select the network structure. However, there are several types of problems where neural network

methods are more applicable than abductive network. For example, abductive network only performs supervised learning where the input and outputs are known. In cases where the classes are not known, unsupervised neural network techniques are more appropriate.

2.4.2 Comparisons to statistical regression

Regression is the process of finding a mathematical function that best represents the relationships among input (independent) and output (dependent) variables in a database. Most regression techniques are parametric; they require the user to specify the functional form of the solution. A basic form of parametric regression is linear regression. This type of regression generally results in inaccurate models unless one knows or happens to guess the correct underlying form of the relationships. Often designers spend significant time experimenting with different functional relationships and regression algorithms to obtain acceptable models. If no assumptions about the type of function or variable distributions are used, the regression is non-parametric. The abductive network synthesis process can be classified as a form of non-parametric regression since it discovers the best network architecture for a given database of examples [31].

2.5 Evaluation System

Our evaluation system first undergoes a training phase which selects connectivity and adjusts the summation weighting functions that used to map the BSDB into the predicted MOS. The output speech sentences that processed by high performance coder (MOS=3.79) and a low performance coder (MOS=2.27) with the input speech sentences are used to prepare the learning data. The learning data base set contains 48 BSDB vectors, each of 13 elements (cover a spectrum from 300 to 3400 Hz that used in telephony), with their desired speech quality scores, each of 11 elements that match the

output layer's size. During the learning phase of our evaluation system, the actual output speech quality scores is compared to the desired scores and the errors between the actual and desired scores are then used to determine the best network structure, element types, coefficients, and connectivity that minimize the predicted square error (PSE) at certain value of the complexity penalty multiplier (CPM). This process is repeated using different values of CPM, varies between 0.35 to 0.6 in step of 0.5. At the end of learning phase, we have some evaluation systems, each achieved mapping of BSDB into the predicted speech quality scores. We tested all of these systems to predict the average MOS of two other coders (high performance coder with MOS=3.73 and low performance coder with MOS=2.26), to choose only one evaluation system that gave the lowest prediction error. Figure 2.5.7and 2.5.8 depict the production phase of our sub-evaluation system in which four impairments were used to create it while Fig. 2.5.9 shows the predicted MOS error versus CPM for the evaluation systems of Fig.2.5.8. To produce only one evaluation system, eliminate the dependency of our evaluation system on the training coders speech sentences, and make sure that we extract the perceptual features from the training speech sentences, we produce three other sub-evaluation systems that used a pair of coder (high and low ratings) for training and other pair for choosing the best sub-evaluation systems. As a result, our evaluation system composed of four sub-evaluation systems and the predicted ratings will be the average output of each sub-evaluation system. Figure 2.5.10 illustrates the abductive network structure for one of the sub-evaluation system where b_i is bark spectral distance for the band i.

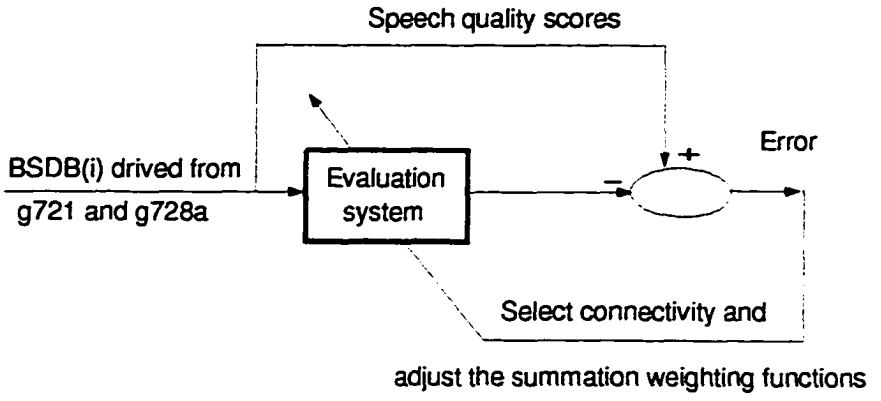


Figure 2.5.7 Learning phase for one of the sub-evaluation system.

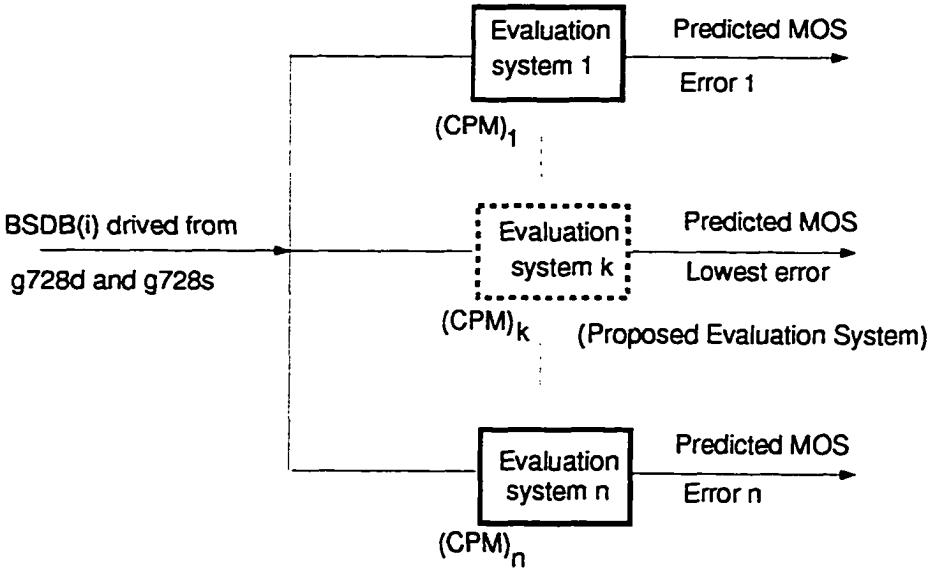


Figure 2.5.8 Selection one of the sub-evaluation system.

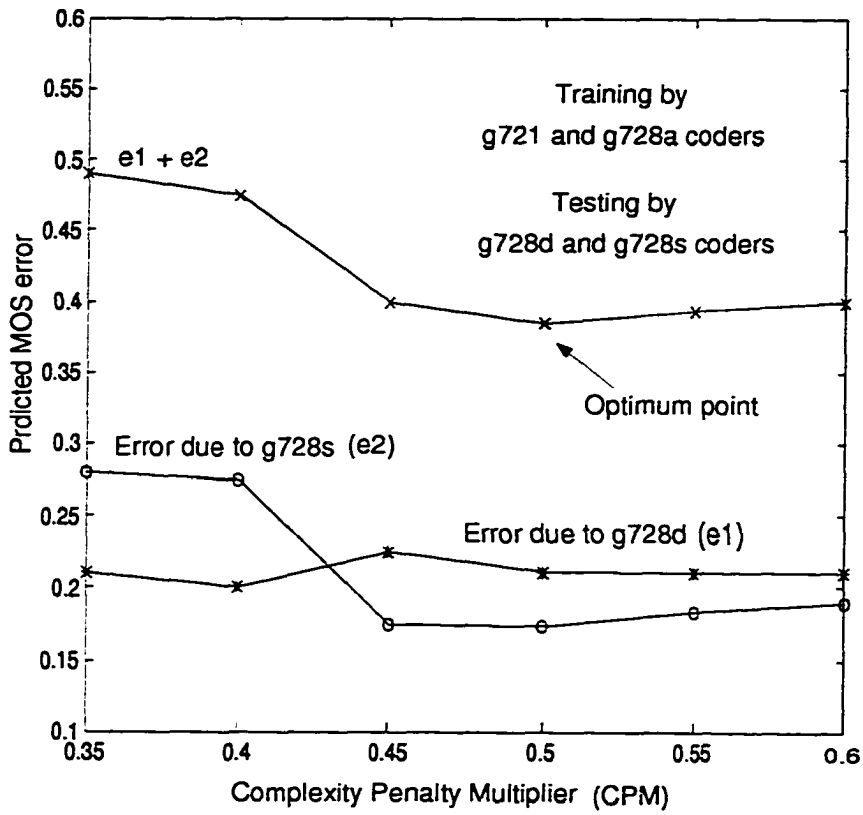


Figure 2.5.9 Predicted MOS error versus CPM for the evaluation system of Fig. 2.5.8.

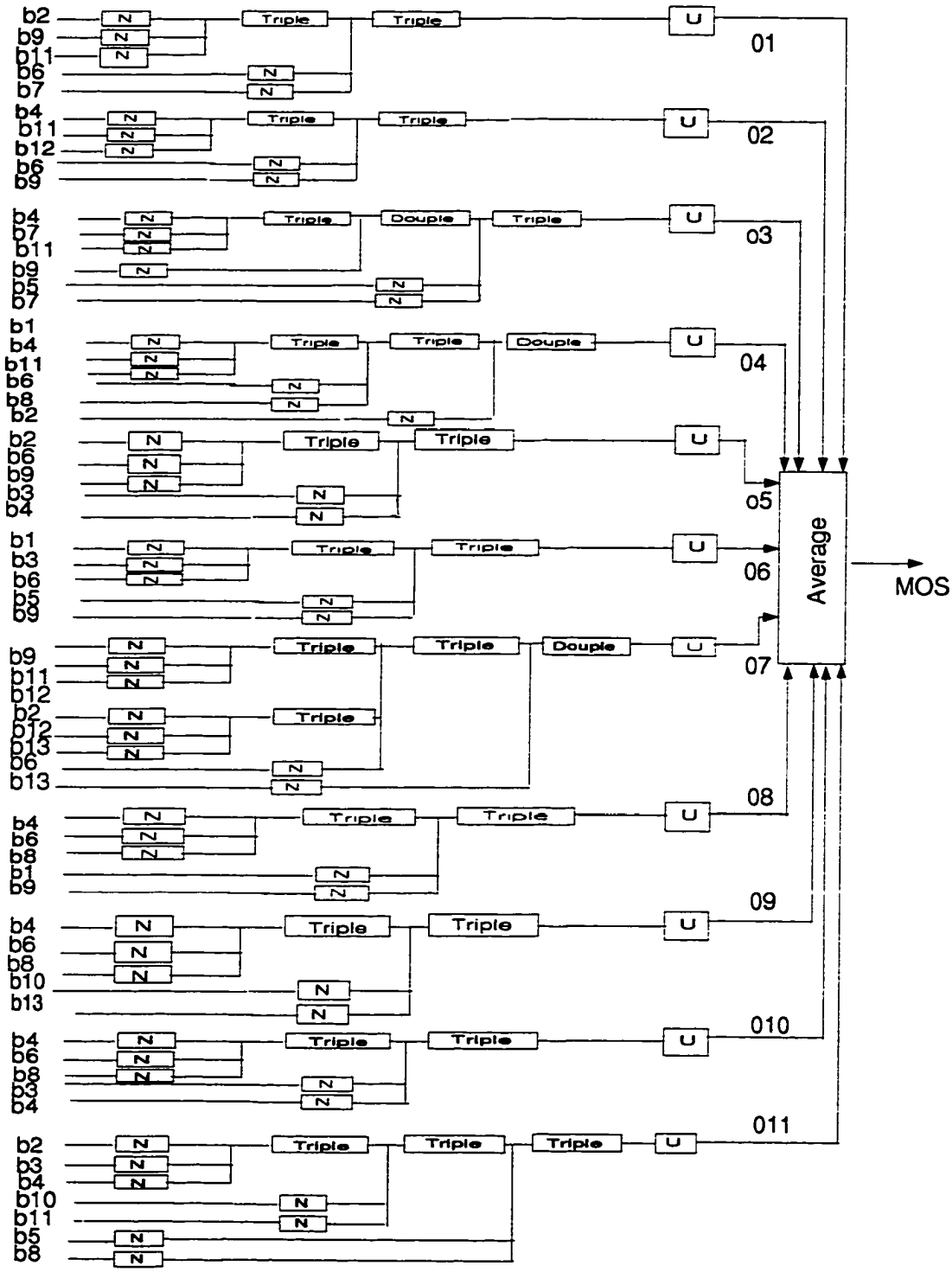


Figure 2.5.10 The abductive network structure of the sub-evaluation system. The symbol b_i is the bark spectral distance per band i.

2.6 Numerical Results

In this section, we demonstrate the validity of the proposed evaluation system by comparing the average predicted MOS obtained from our technique to those obtained from the subjective test. The symbol r, that will appear in the following figures, is the correlation coefficient between the actual and predicted MOS values while the symbol s is the standard deviation of the prediction error. For Fig. 2.6.11-2.6.13, we used the speech files that processed by four coders, in training our evaluation system. Then we use it to predict the performance of the other coders. The actual MOS and the predicted one for mixed speakers is illustrated in Fig. 2.6.11 while MOS ratings, based on female/male speakers, are shown in Fig. 2.6.12/Fig. 2.6.13.

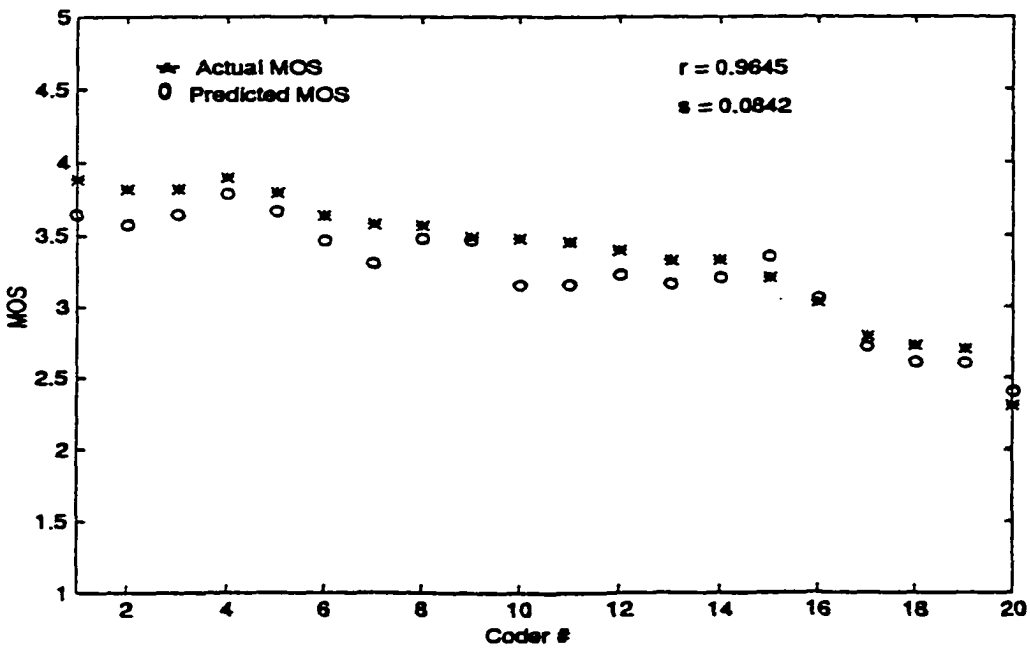


Figure 2.6.11 Actual and predicted MOS values for mixed speakers.

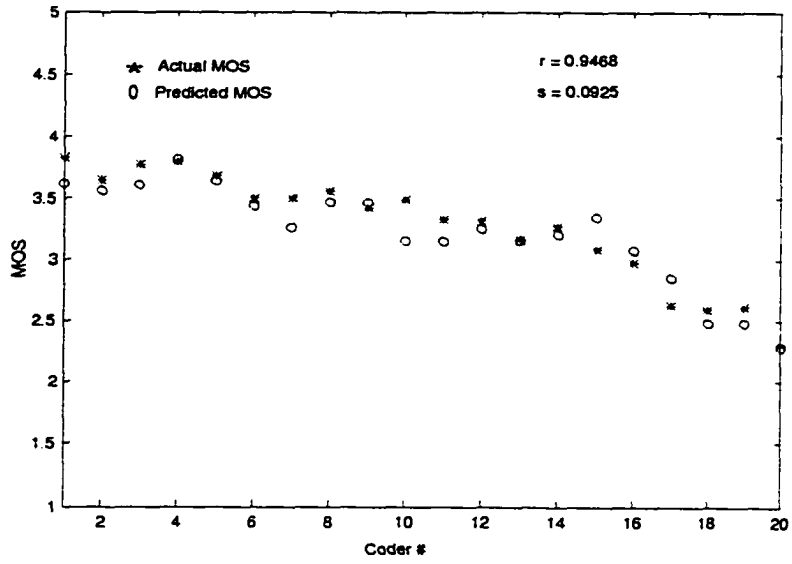


Figure 2.6.12 Actual and predicted MOS values for female speakers.

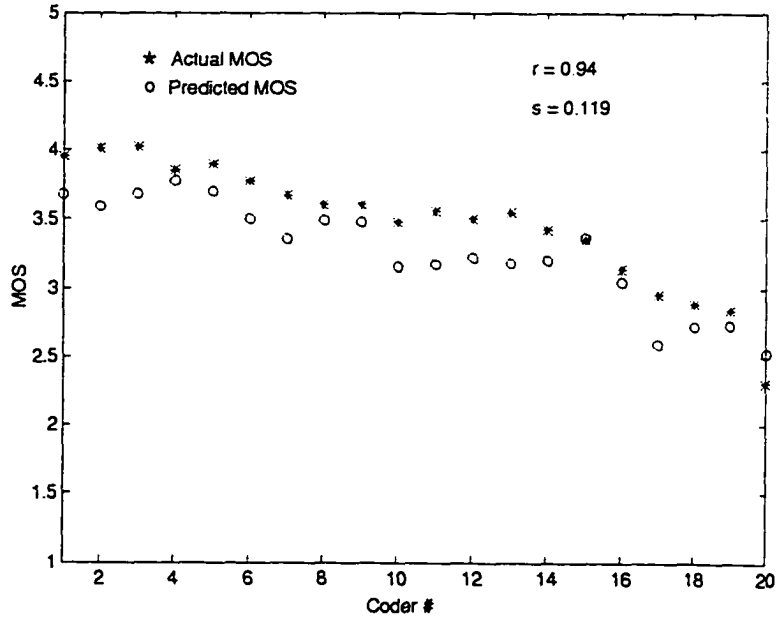


Figure 2.6.13 Actual and predicted MOS values for male speakers.

It is clear that our evaluation system is robust in evaluating the MOS ratings. For example, the actual MOS for coders #16 (bit rate = 7.95 Kbps), coder #21 (bit rate = 4.8 Kbps), and coder #5 (bit rate = 16 Kbps) are 3.49, 3.03, and 2.31 respectively and the predicted MOS for them are 3.465, 3.058, and 2.4 respectively.

To ensure that our technique is reliable, we prepared another evaluation system using the speech files that processed by different combination of four coders (two high performance coders, two low performance coders). The results of using the new evaluation system in predicting the MOS ratings are depicted in Fig.2.6.14–2.6.16.

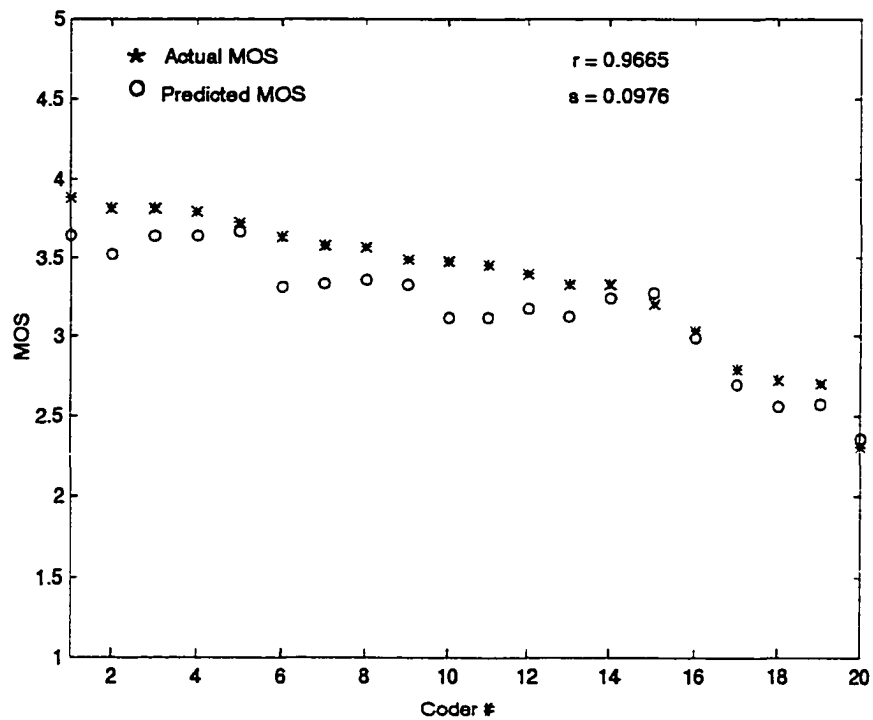


Figure 2.6.14 Actual and predicted MOS values for mixed speakers.

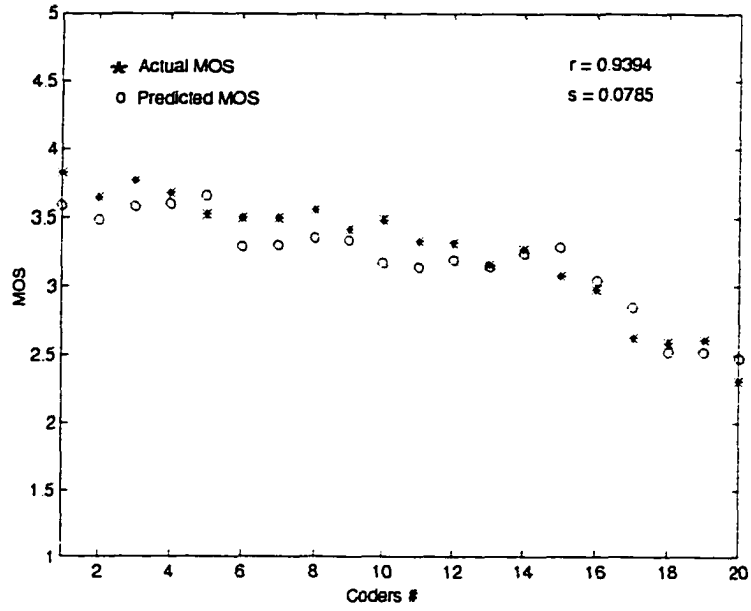


Figure 2.6.15 Actual and predicted MOS values for female speakers.

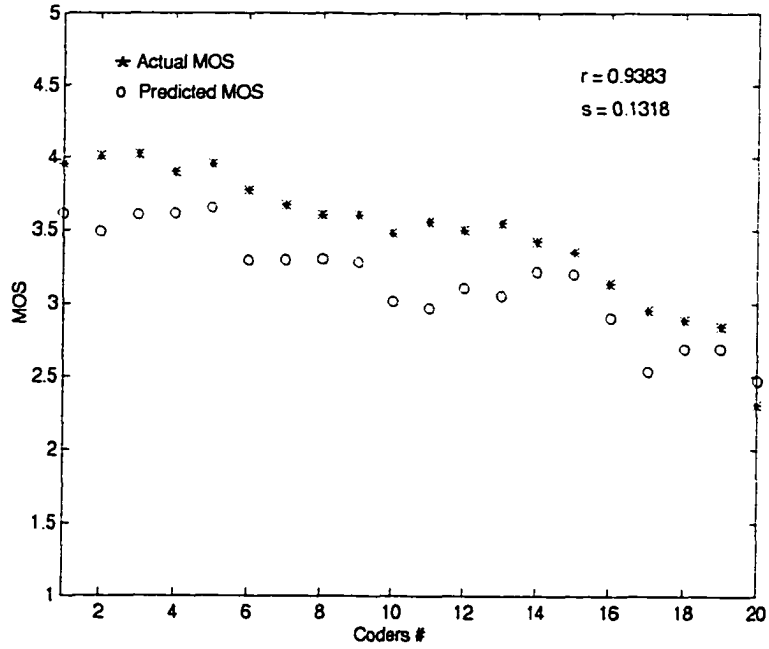


Figure 2.6.16 Actual and predicted MOS values for male speakers.

From Fig. 2.6.14, the predicted MOS value for the coder #16, coder #21, and coder #5 are 3.38, 2.988, and 2.35 respectively. This results explains that our technique is reliable in evaluating the coders' performance.

The study of the figures shows that:

• Our technique successfully predicts speech quality that are highly correlated to human responses across a wide range of quality levels and coding algorithms,

• Our technique performs best for mixed speakers followed, with a fair judgement, by female or male speakers,

After performing extensive experimentations, Table 2.6.2 illustrates some results that show:

• Sensitivity of the prediction performance of the proposed evaluation system to the preprocessing operations of the speech database (conditions 1–3),

• Using back-propagation neural network , instead of the abductive network in predicting the speech quality (condition 4),

• Sensitivity of the evaluation system performance to the training data set's length (condition 5).

#	Conditions	r	s
1	Without removing silent period, and without applying average equivalent power process between i/p and o/p speech records.	0.9237	0.1088
2	Removing silent period, and without applying average equivalent power process between i/p and o/p speech records.	0.8947	0.1567
3	Removing silent period, and applying average equivalent power process between i/p and o/p speech records.	0.9452	0.0789
4	Applying condition # 3 and using back-propagation neural network.	0.718	0.2166
5	Applying condition 3 and using 96 BSDB vectors ,instead of 48, for training the abductive network.	0.9131	0.138

Table 2.6.2 Sensitivity of the prediction performance to the preprocessing operations, training set, and neural networks.

2.7 Validity of the evaluation system concept with conventionally-oriented parameters

In this section, we use the speech records that have a flat response from 0 to 4 kHz. Therefore, we have 17 critical band filters. Following the previous concepts (section II) we will use the average signal-to-distortion ratio (SDR), instead of BSDB, to introduce an evaluation system that map the classical-oriented parameters (SDR) into estimated MOS. According to [33], the signal-to-distortion ratio (SDR), denoted QS(i) is given by:

$$QS(i) = 10log_{10} \frac{\sum_{j \in b_i} | X(f_j) |^2}{| \sum_{j \in b_i} | X(f_j) |^2 - \sum_{j \in b_i} | Y(f_j) |^2 |} \qquad (2.7.10)$$

Where j ranges over all frequencies specified for the i-th band, b_i, X(f) and Y(f) are discrete Fourier transforms of a given input and output speech frame. The average SDR is the average of QS(i) over all frames. Calculating the critical frequency bands b_i that used in eq. 2.7.10 has been done by two methods:

2.7.1 Method 1

Following the same procedure of section I with cancelling the effect of equal loudness and intensity compression process, the weighting functions (that preserve the perceptual effect, Fig. 2.3.2.3 are converted to the critical band filters that shown in Fig. 2.7.1.17.

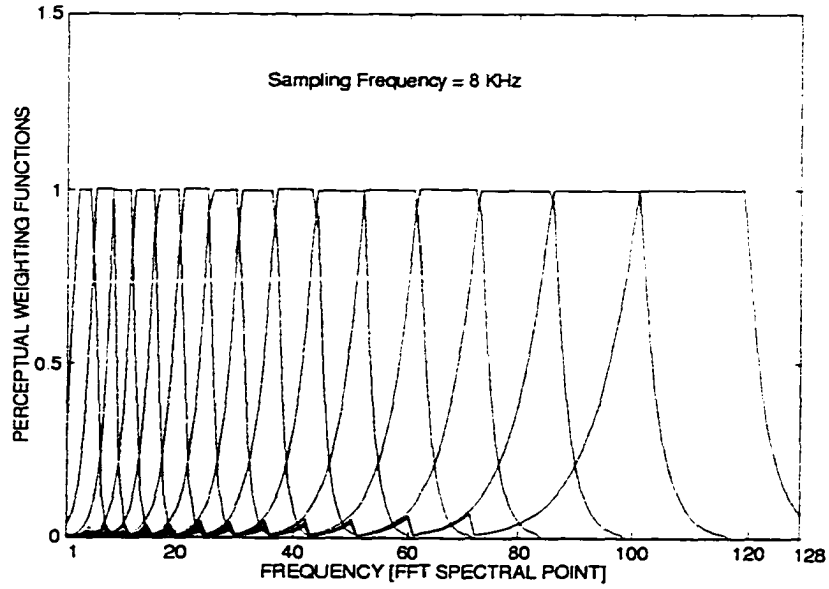


Figure 2.7.1.17 Critical band filters used in calculation the average signal-to-distortion ratio.

2.7.2 Method 2

In this method, we would like to examine our evaluation system when we use SDR that are calculated by a very simple preprocessing operations. In this case, we use 5 ms non-overlapping frames, no hamming window, and we divide the spectrum (0–4 kHz) into 18 flat critical bands using the same critical filters' bandwidths that appeared in [34] and listed in Table2.7.2.3.

Band	From	To	Bandwidth
1	0	100	100
2	100	200	100
3	200	300	100
4	300	400	100
5	400	510	110
6	510	630	120
7	630	770	140
8	770	920	150
9	920	1080	160
10	1080	1270	190
11	1270	1480	210
12	1480	1720	240
13	1720	2000	280
14	2000	2320	320
15	2320	2700	380
16	2700	3150	450
17	3150	3700	550
18	3700	4400	700

Table 2.7.2.3 Filters' bandwidths.

Table 2.7.2.4 compares the results, for only 8 different coders, using BSDB (17 samples that cover 0–4 kHz spectrum) and the average signal-to-distortion ratio derived from the previous two methods.

Coder#	Bit-rate Kb/s	Remark	actual MOS	Predicted MOS		
				BSDB	Average SDR Method 1	Average SDR Method 2
1	32		3.68	3.686	2.67	3.1484
2	32	With silence coding	3.73	3.685	2.716	3.056
6	16	10 % fe rate	2.99	2.82	3.052	3.116
7	16	3 % fe rate	2.85	2.827	3.08	2.973
10	16	with silence coding	3.7	3.59	3.709	3.7455
11	16		3.76	3.72	3.178	3.624
13	13	fixed point	3.56	3.27	3.36	3.85
21	4.8	floating point	3.06	2.92	2.84	2.95

Table 2.7.2.4 Actual and predicted MOS values for different objective evaluation techniques

It is clear that the evaluation system that used the perceptually parameters is much reliable than the other, especially in predicting the performance of the coders that have bit rate different than that used in the training phase. We may say that the evaluation system in this case learned the perceptual aspects of the speech distortion that are not change too much with bit rate variation. But, the evaluation system that used the classical parameters (SDR) predict well the performance of coders that have the same bit rate of the training phase. In this case, the system learned the electrical aspects of the speech distortion that doesn't change too much for the same bit rate coders.

The above results indicate that the evaluation system performs best with the perceptually oriented parameters.

2.8 Conclusion

In this chapter, we introduced a perceptually-oriented objective measure technique in which we use the abductive networks, that evolved from neural network, statistical modeling, and artificial intelligence concepts to map the bark spectral distortion per band (BSDB) into estimated speech quality. The results indicate that our proposed technique is reliable and robust in evaluating the coded speech quality and it is highly correlated to human responses across a wide range of quality levels and for a wide range of speech coding techniques.

Chapter 3 USING RADIAL BASIS FUNCTION NEURAL NETWORK IN PREDICTING THE SPEECH QUALITY

3.1 Overview

The purpose of this chapter is to provide another perceptually-oriented objective measure technique that is highly correlated to human responses across a wide range of quality levels and for a wide range of speech processing, transmission, and transport technologies [13]. In this chapter, we introduce our objective measure technique that:

• Emulates several known features of perceptual processing of speech sounds by human ear (including critical-band masking, equal loudness, and the intensity-loudness power law operations) to map the speech power spectrum into auditory power spectrum (bark domain),

• Derives the perceptual LPC coefficients from the auditory spectrum that used to calculate, for each frame, the cepstrum distance between the input and the output coded speech signals,

• Uses the radial basis functions neural network to map the perceptual cepstrum distance per frame into the corresponding estimated speech quality.

Parameters derived from speech materials of ten different coders are used in the training phase of the proposed neural network. Validation of our technique is proved by using the trained neural network in estimating the performance of other sixteen different speech coders. The results indicate that our proposed technique is reliable and robust in evaluating the coded speech quality. A major achievement of our objective measure

technique is the realization of a speech evaluation method that is technology independent (independent of coding algorithm and communication technology).

The rest of this chapter is organized as follows: section 2 introduces the calculation details of the perceptual cepstrum distance per frame. Third section shows the radial function neural network concepts. The proposed evaluation system is illustrated in section 4. Section 5 summarizes the numerical results while conclusions and the further work are introduced in section 6.

3.2 Calculation of the Perceptual Cepstrum Distance per Frame (PCDF)

3.2.1 Preprocessing

As a first step, the input and output records are aligned in time. The relative delays are determined by cross correlating the input and output speech envelops. To avoid system gain effects, the records are corrected to have equivalent average power. The speech frame is weighted by Hamming window and the consecutive frames overlapped by 50 %. By computing the magnitude square FFT spectrum, the frame's power spectrum P(f) is calculated and followed by some speech processing operations similar to speech processing that performed chapter 2. These operations are summarized as follows:

- Compute the FFT-based sample spectrum,
- Warp the frequency axis into a Bark frequency scale,
- Apply a critical-band masking curve, i.e., convolve the Bark spectrum with a critical-band filter and sample at 1-Bark intervals,
- Weight the samples for equal loudness to adjust for the ear's frequency dependent loudness sensitivity,

• Apply 1/3 power law to simulate the non-linear relation between sound intensity and perceived loudness.

After performing the above operations, we obtained thirteen samples of the auditory spectrum, B(i), that covers the telephony spectrum (300-3400 Hz). Then, we applied some speech signal operations on B(i) as described in the following section.

3.2.2 Perceptual Cepstrum Distance per Frame (PCDF)

In this step, the inverse DFT is applied to B(i) to yield the autocorrelation function, $r_{ij}(r_j)$, dual to B(i). The perceptual LPC coefficients, α_i, are obtained by solving the Yule-Waker equation, $\sum_{i=1}^{p} \alpha_i \, r_{ij} = r_j$, with p is the order of LPC analysis (we use p=8, 10 , and 12). As a final operation, the perceptual cepstral coefficient, c_n, can be computed recursively from the perceptual LPC coefficients, α_i, by [35]:

$$c_n = -\alpha_n - \sum_{i=1}^{n-1} \frac{n-i}{n} \alpha_i \, c_{n-i} \, , \qquad n \geq 1 \tag{3.2.2.1}$$

Where $\alpha_i = 0$ when i > p. For each frame, l, the perceptual cepstrum distance per frame (PCDF) is defined as [7]:

$$PCDF(l) = \frac{10}{log_e 10} \sqrt{2 \sum_{i=1}^{p} \{c_x(i) - c_y(i)\}^2} \tag{3.2.2.2}$$

Where $c_x(i)$ and $c_y(i)$ are the i-th cepstral coefficients of the input and the output coded speech signals. Figure 3.2.2.1 illustrates the basic transformations used in obtaining the perceptual cepstrum distance per frame PCDF(l).

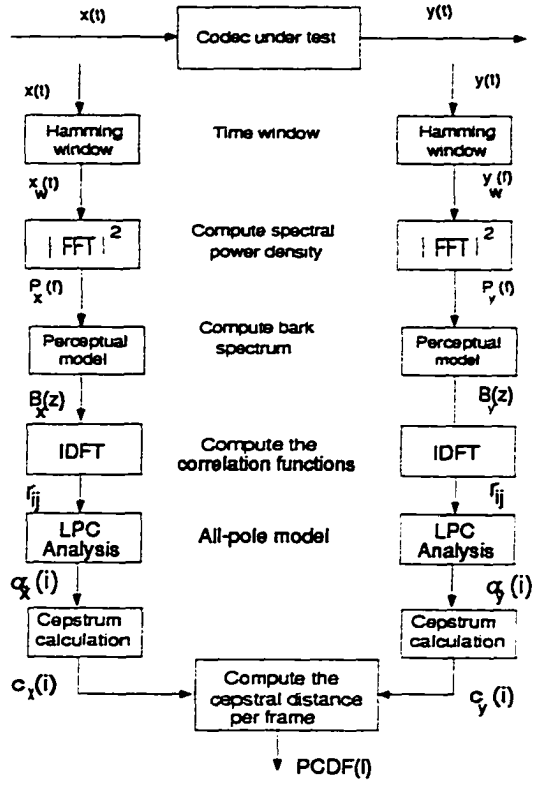


Figure 3.2.2.1 Basic transformations used in obtaining PCDF(l).

3.3 Radial Basis Function Neural Network Concepts

3.3.1 Background

The radial basis function network (RBFN) offers a viable alternative to the two-layer neural network in many applications of signal processing [36]. Like networks with sigmoidal nonlinearities, RBFN's can have the capability to represent arbitrary functions [37], but RBFN's can be trained more rapidly [38]. The RBFN consists of three layers (input, hidden, and output) of nodes with successive layers exhaustively interconnected by feedforward arcs [39]. The input data vector is presented via the input layer which fans out the input data without making calculations. The data flows along the connections toward the hidden layer that performs a nonlinear transformation and maps the input space

into a new space. The output layer then implements a linear combiner on this new space to produce the output vector. The transfer functions of the hidden nodes are given by [39]:

$$a_h = exp\left(- \| x - m_h \|^2 / \sigma_h^2 \right), \quad h = 1..... H \qquad (3.3.1.3)$$

Where a_h is the output unit h in the hidden layer, of size H, given the input x, m_h is the position of the center (weight) of the radial unit in the input space , and σ_h is the unit width. Each RBF unit has a significant activation over a specific region determined by m_h and σ_h.

3.3.2 Network simulation

RBF neural networks were simulated using Matlab software package[40]. During the learning phase of RBF network, number of the hidden neurons is created one a time. At each iteration, the input vector which will result in lowering the network error, the most, is used to create the neuron's weight (center) vector. The neuron's bias is adjusted by choosing a suitable value of a spread constant, sc. This constant determines how wide the radial basis functions are. The weights and biases of the output layer are calculated by linear least square method in which the sum-squared error between the desired and the actual output is minimized. The error of the network is checked, and if low enough the design steps are finished. Otherwise the next neuron is added. This procedure is repeated until the error goal is met, or the maximum number of neurons is reached. For our technique, the network has 238 neurons in the hidden layer and 11 neurons in the output layer. The only real design decision for the radial basis networks (besides picking an error goal) is finding a good value for the spread constant. It is important that the radial functions of the hidden layer overlap so as to allow good generalization. However,

the radial basis functions should not be so spread out such that the radial basis neurons return outputs near 1 for all the input vectors used in design [40].

3.4 Evaluation System

3.4.1 Preparing the learning data set

The speech material consists of sentence pairs with approximately 0.4 sec (3200 samples) silent period at the beginning of each sentence. Figure 3.4.1.2 shows one of our speech file record. In our technique, we eliminate the effect of the silent period at the beginning of each sentence by dividing the speech record into two parts. Each part has one sentence denoted by S1 and S2. Then we remove most of the silent period at the beginning of each sentence as described in Fig. 3.4.1.3 illustrate that concept. In our algorithm, the total evaluation system has two sub-evaluation systems, each assigned for one sentence, and the estimated speech quality score is the average of these sub-evaluation scores.

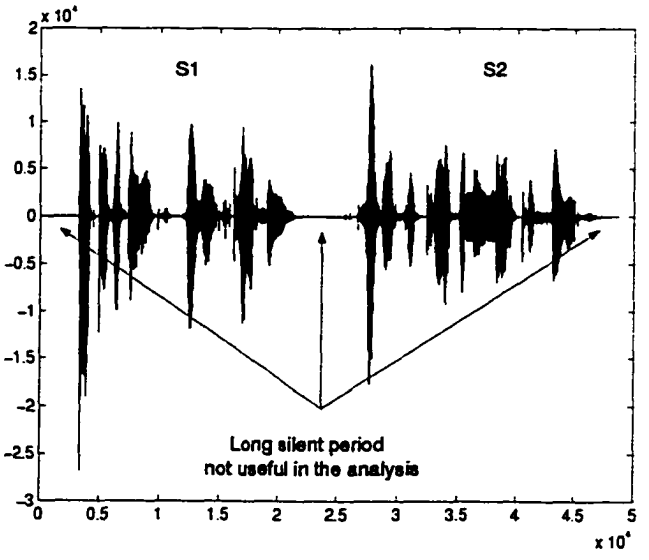
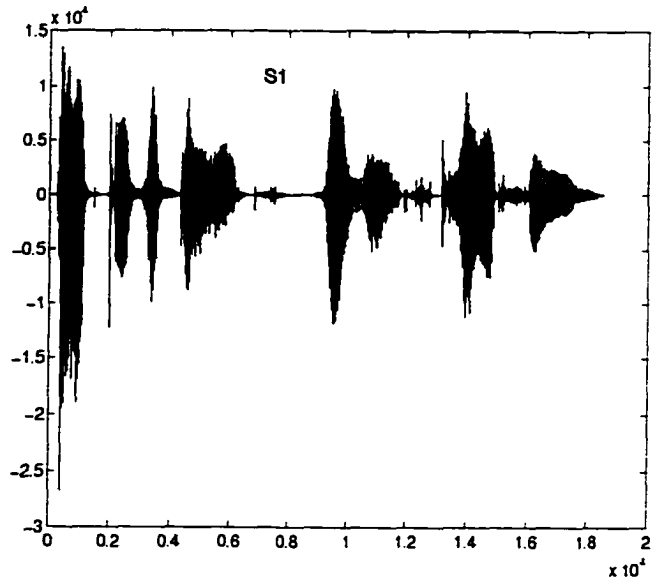
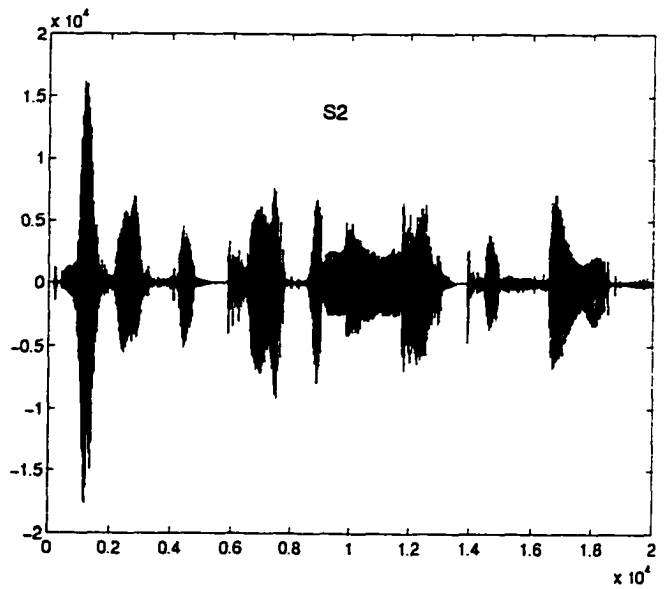


Figure 3.4.1.2 One of our speech files. S1 and S2 are denoted to the first and the second sentences respectively



(a)



(b)

Figure 3.4.1.3 Removing the silent period at the beginning of the two sentences: (a) sentence one, (b) sentence two.

To have a constant feeding input to the Neural Network, we divide each sentence into fixed number of frames. Typically we use 211 overlapping frames with their sizes in the range of 20 to 30 msec. We believe that dividing the sentence into fixed number of frames with different size and then approximate that frame by all-pole model, with a suitable order, is more accurate than using the time warping technique that modifies the time shape of the speech signal.

The output speech sentence (S1 and S2) that processed by 10 different coders with the input speech sentence (S1 and S2) are used to prepare the learning data for the two sub-evaluation systems. The learning data set contains 240 vectors (24 speech files for each coder) each vector has 211 elements (PCDF). Associated with the input vectors their desired speech quality scores, each of 11 elements that match the output layer's size. It is obvious that increasing number of frames will increase the dimension of each input vector and then we will need more coders to prepare the learning data and increase the learning time. In the other hand, decreasing the number of frames will increase the frame size and that yields to decrease the result accuracy.

3.4.2 Neural network's learning phase

During the learning phase of our evaluation systems, the actual output quality scores is compared to the desired scores and the errors between the actual and desired scores are then used to determine the best network coefficients that minimize the predicted square error (PSE) for certain value of spread constant, sc. The learning procedures is repeated for different values of sc. Figure 3.4.2.4 shows the PSE during one of the learning phase while Fig. 3.4.2.5 depicts the production phase of the sub-evaluation system.

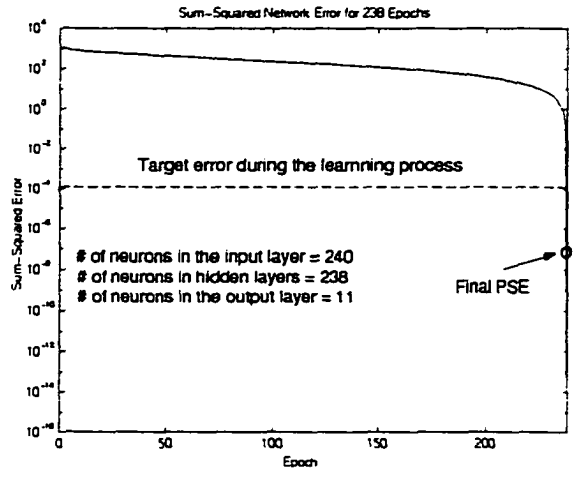


Figure 3.4.2.4 Predicted square error for the learning phase of the radial basis functions neural network.

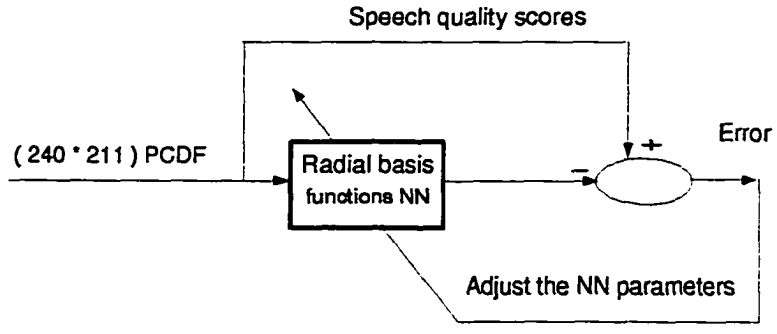


Figure 3.4.2.5 Learning phase for one of the sub-evaluation system.

3.4.3 Neural Network's test phase

At the end of the learning phase, we have some evaluation system, each achieved mapping of PCDF into the predicted speech quality scores. We tested all of these systems to predict the average MOS of other 16 coders. We chose the one that gave the most accurate result for the two sub-evaluation systems. So our evaluation system consists of the best sub-evaluation systems. This evaluation system can be used to predict the performance of speech communication systems including speech coders and mobile communication systems.

3.5 Numerical Results

In this section (with using the order of the perceptual LPC analysis, p, equals to 10) we demonstrate the validity of the proposed evaluation system by comparing the average predicted MOS obtained from our technique to those obtained from the subjective test. The symbol r, that will appear in the following figures, is the correlation coefficient between the actual and predicted MOS values while the symbol s is the standard deviation of the prediction error. For Fig. 3.5.6–3.5.10, we used the speech files that processed by ten coders (Group 1) in training our evaluation system. Then we use it to predict the performance of the sixteen other coders. The actual MOS and the predicted one, based on the mixed speakers, is illustrated in Fig. 3.5.6 and Fig. 3.5.7 for the first and the second sub-evaluation system respectively.

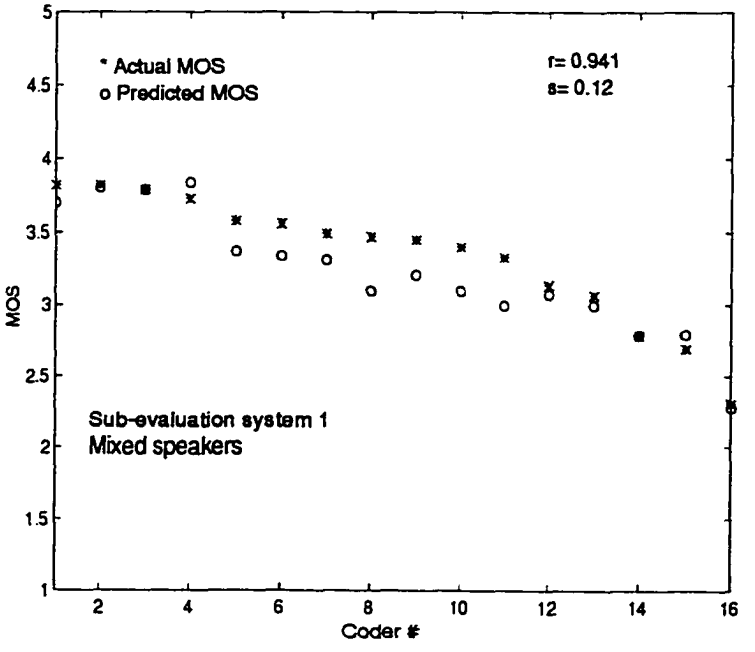


Figure 3.5.6 Actual and predicted MOS values for sub-evaluation system 1.

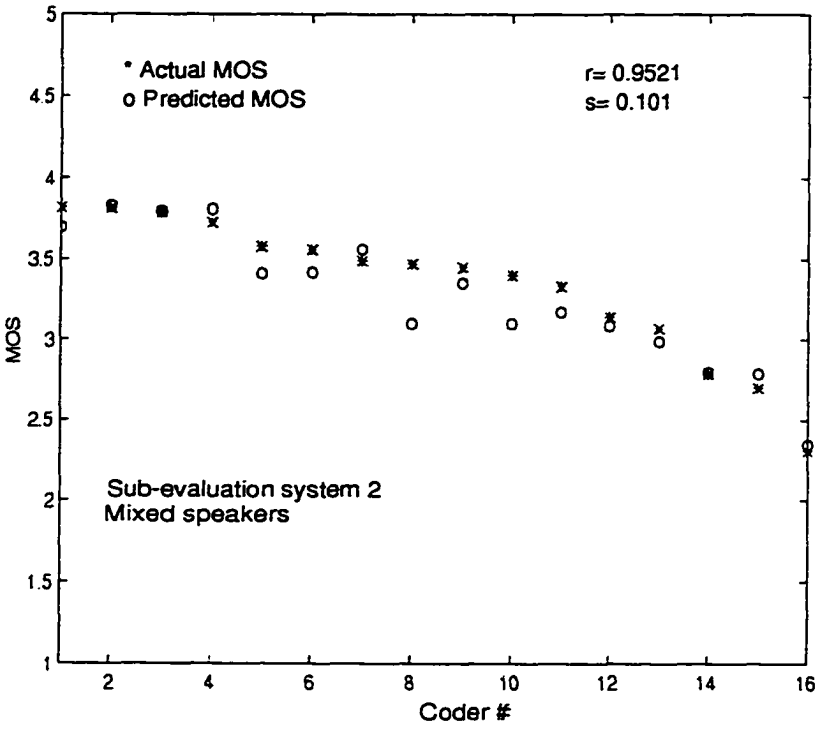


Figure 3.5.7 Actual and predicted MOS values for sub-evaluation system 2.

As we mentioned, the net result of our technique is the average score of the two sub-evaluation systems. Fig 3.5.8 compares the actual and the predicted MOS based, on the mixed speakers, while Fig. 3.5.9/Fig. 3.5.10 illustrate the results based on female/male speakers respectively.

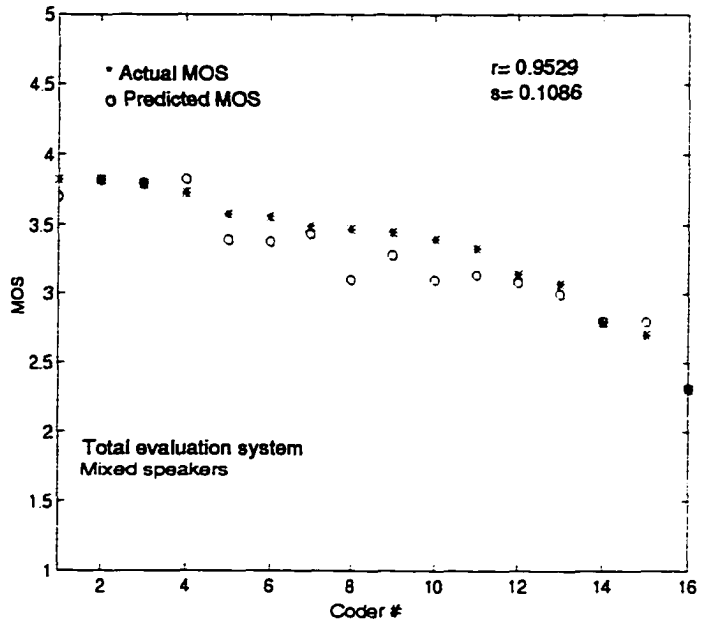


Figure 3.5.8 Actual and predicted MOS values based on mixed speakers for the total evaluation system.

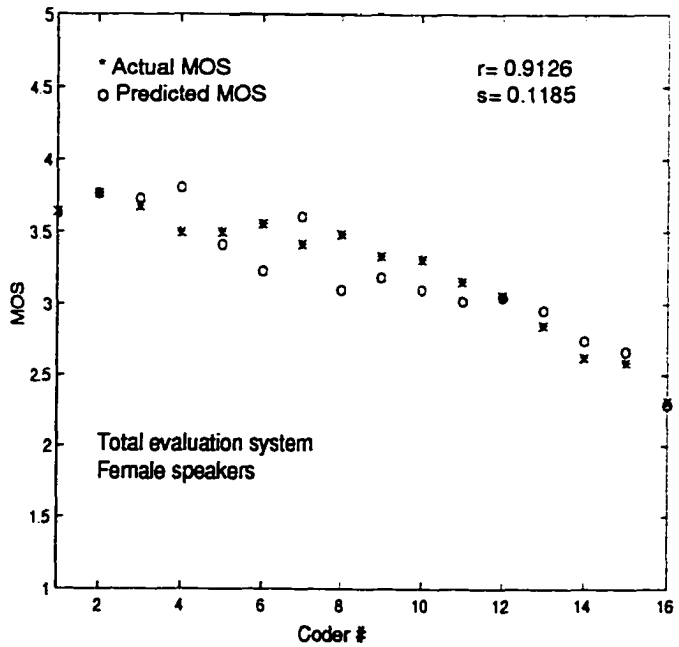


Figure 3.5.9 Actual and predicted MOS values based on female speakers for the total evaluation system.

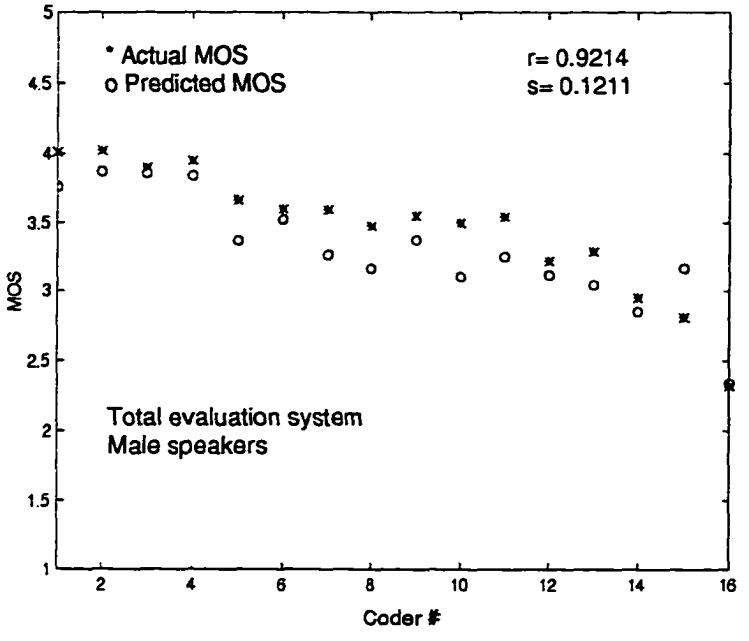


Figure 3.5.10 Actual and predicted MOS values based on male speakers for the total evaluation system.

Figure 3.5.8 shows that our evaluation system is robust in evaluating the MOS score. For example, the actual MOS for coders "g728r14" and "g728v" (both has a bit rate of 16 kbps) is 3.82 while the predicted MOS for them are 3.704 and 3.8 respectively. In the other hand, the maximum predicting error of approximately 0.33 in rating the MOS score of the two coders "drod" and 'cplus68" (both has a bit rate of 6.25 kbps). To ensure that our technique is reliable , we prepared another evaluation system using the speech files that processed by different mixed coders (Group 2). The results of using the new sub-evaluation systems in predicting the MOS ratings are depicted in Fig. 3.5.11-3.5.12. Figure 3.5.11 depicts the evaluation performance for the first sub-evaluation performance while Fig. 3.5.12 introduces the result for the second sub-evaluation system.

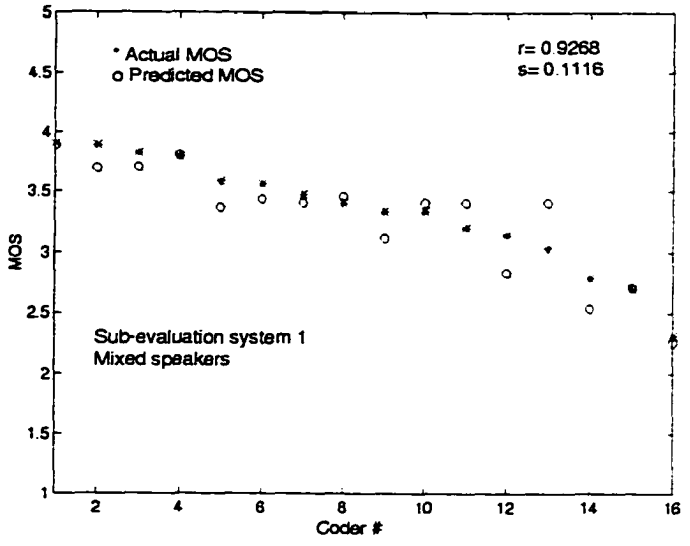


Figure 3.5.11 Actual and predicted MOS values for sub-evaluation system 1.

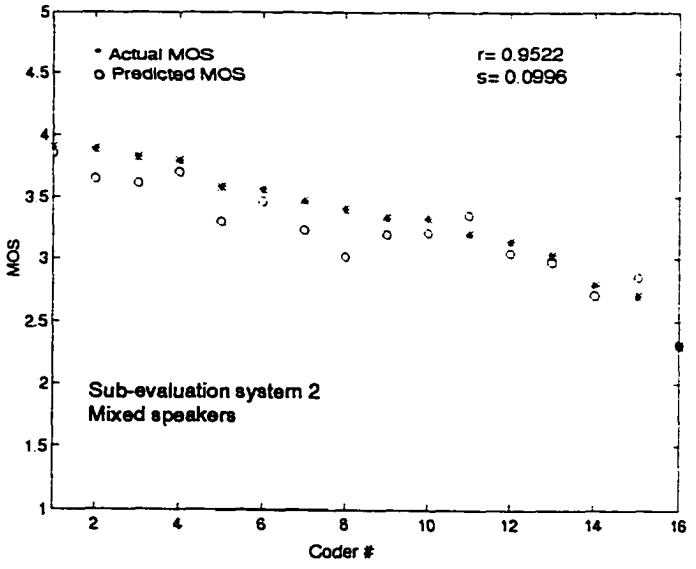


Figure 3.5.12 Actual and predicted MOS values for sub-evaluation system 2.

Figures 3.5.13-3.5.15 introduce the net result of the predicting performance of the new evaluation system based on the mixed, female, and male speakers respectively.

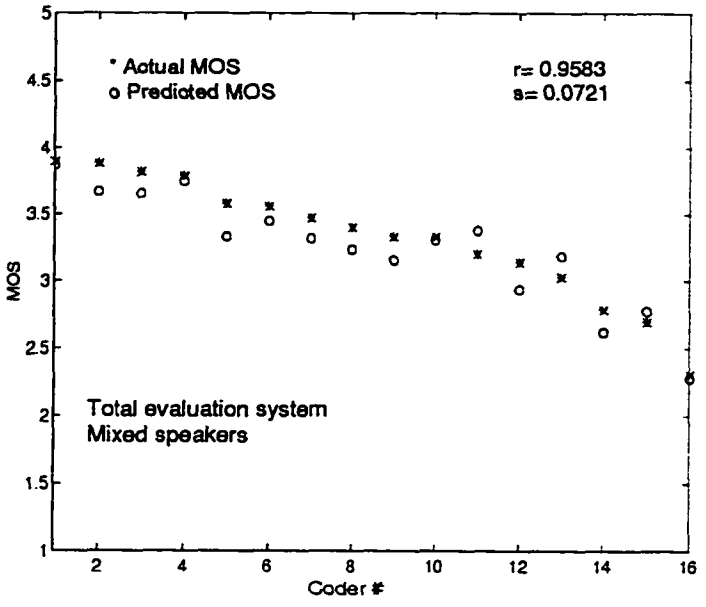


Figure 3.5.13 Actual and predicted MOS values based on mixed speakers for the total evaluation system.

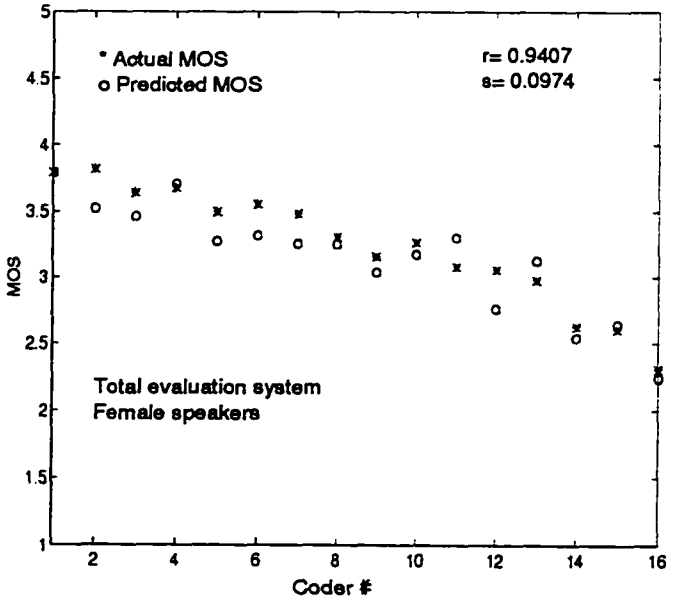


Figure 3.5.14 Actual and predicted MOS values based on female speakers for the total evaluation system.

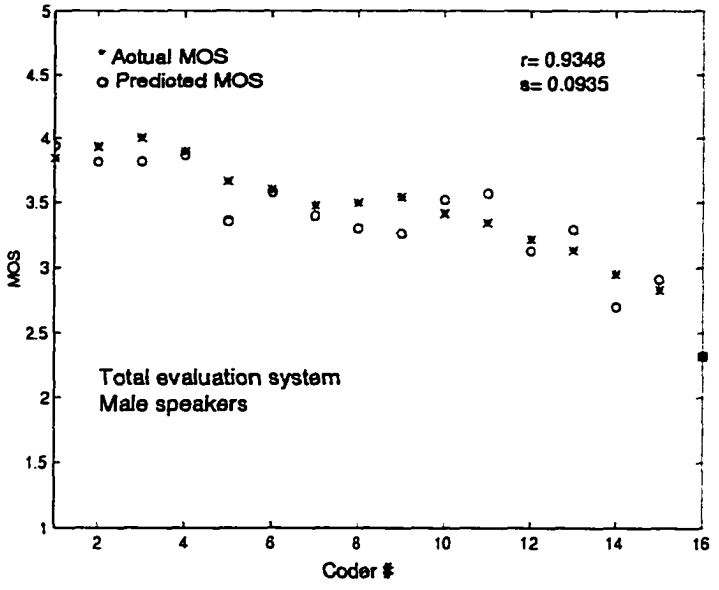


Figure 3.5.15 Actual and predicted MOS values based on male speakers for the total evaluation system.

From Fig.3.5.13, we note that choosing another learning data set yielded to decrease the maximum predicting error to approximately 0.27 instead of 0.33 as mentioned in Fig. 3.5.8. The predicting performance of the proposed technique can be improved by choosing appropriate learning data set. We repeat our algorithm using different values of p (order of the perceptual LPC analysis). Table 3.5.1 introduces the predicting accuracy result and it is clear that p=10 gives better estimating performance than the other two values.

Order of the perceptual LPC analysis, p	Correlation coeffecient, r	Standard deviation of the predicting error, s
8	0.89	0.21
10	0.9583	0.0721
12	0.9	0.157

Table 3.5.1 Effect of the order of the LPC analysis in the predicting performance.

The study of the figures shows that:

• Our technique successfully predicts speech quality that are highly correlated to human responses across a wide range of quality levels and coding algorithms,

• Our technique performs best for mixed speakers followed, with a fair judgement, by female or male speakers,

• Predicting performance of sub-evaluation system 2 is better than of the first sub-evaluation system. This indicates that the MOS score of the second sentence has a stronger effect on evaluating the whole speech record than that of the first sentence has,

• The predicting performance of the proposed technique can be improved by choosing appropriate learning data set.

3.6 Conclusion

In this chapter, we introduced a perceptually-oriented objective measure technique in which we use the radial basis function neural network to map the perceptual cepstrum distance per frame (PCDF) into estimated speech quality. The results indicate that our proposed technique is reliable and robust in evaluating the coded speech quality and it is highly correlated to human responses across a wide range of quality levels and for a wide range of speech coding techniques. As a further work we are trying to evaluate a new algorithm in which we extract some features from the output coded speech signals, without referring to the input speech signal. And using the radial basis function neural network to estimate the speech quality.

Chapter 4 DEGRADATION EFFECT OF CELL LOSS ON SPEECH QUALITY OVER ATM NETWORK

4.1 Overview

Broadband Integrated Service Digital Network (B-ISDN) will transport diverse classes of traffic such as data, voice, image, and video. ATM (Asynchronous Transfer Mode) is being standardized as the transport mechanism to integrate such services in a single network [41]. These services are likely to have a wide range of traffic characteristics, performance, and quality of service (QOS) requirements [42]. ATM poses some problems when applied to transmission of real-time sources such as speech [43]. Among the central problems in the support of real-time applications (voice, video) with ATM networks are the existence of delay jitter [44] and cell loss. Designers of speech coders and networks need to work separately and together to heighten our understanding of QOS as perceived by the user [1]. The need for a pre-connection quality of service for statistically multiplexed connections must be assessed [42].

In this chapter, our objective is to understand the impact of cell loss on the speech quality over ATM networks. Understanding of that impact is important for the proper design of network algorithms such as routing, flow control, and management techniques. The management techniques achieve the objective of maintaining the QOS of the ATM layer by managing the number of connections that are accepted and assigning prioritizing to control the jitter and cell loss tolerances. In emerging technology, the user expects a minimum guaranteed value of QOS regardless of traffic intensity, service variety, or network imperfections [45]. A careful definition of the user requirements would also

greatly assist in the design of future telecommunication systems, services [1], and audio applications [46].

The objective measure technique that introduced in chapter 2 is used to study the degradation effect of the transmission of speech over ATM networks. The validity of that measure technique has been checked, as seen in chapter 2, and has been found that it is highly correlated to human responses across a wide range of quality levels and for a wide range of speech processing, transmission, and transport technologies. In that algorithm, as illustrated in chapter 2, we emulate several known features of perceptual processing of speech sounds by human ear (including critical-band masking, equal loudness, and the intensity-loudness power law operations) to map the speech power spectrum into auditory power spectrum (bark domain). Then, we use the auditory power spectrum in calculating the bark spectral distance per band (BSDB) between the input and the output speech signals. Finally, we use the abductive networks, that evolved from neural network, statistical modeling, and artificial intelligence concepts, to estimate the speech quality from the BSDB.

In the following section, section 2, we study the impact of cell loss on the speech quality over ATM networks. Moreover, we compare the results between two replacement techniques: stuffing silent samples and inserting the previous information in the lost cell. In section 3, conclusion is presented.

4.2 Impact of Cell Loss on Speech Quality

Each cell generated by a source is routed to the destination via a sequence of intermediate nodes. Cells may be rejected at the intermediate nodes because of buffer overflow or the delay of that cell goes behind a pre-define upper delay limit, used in reconstruction of

the speech cells. When a cell is lost, the receiver coder needs to deal with the resulting discontinuity in the output signal in some way. Stuffing zero samples in the lost slot or replacing the information from a previous unlost slot are two simple possibilities [47].

For certain loss-rate distribution (uniform, binomial, and Poisson), and speech signal, we define the number of the losing cells and replace these cells either by a silent samples or by the samples of the previous cell. We use the abductive network to predict the speech quality of speech files (24 files) for certain bit-rate (coder algorithm) by feeding the trained abductive network with the BSDB between the corrupted output speech (output of ATM network) and the input speech signal (input of ATM network).

4.2.1 Numerical result

In this section, we illustrate the impact of cell loss-rate (up to 10%) on the speech quality over ATM networks. Moreover, we compare the results between two replacement techniques: stuffing silent samples and inserting the previous information in the lost cell.

In this section, we illustrate the impact of cell loss-rate (up to 10%) on the speech quality over ATM networks. Moreover, we compare the results between two replacement techniques: stuffing silent samples and inserting the previous information in the lost cell.

4.2.1.1 Replacement of the lost cell by silent samples

Figures 4.2.1.1–4.2.1.3 show the relation between speech quality and cell loss for speech bit rate of 128, 32, 16, 13, and 8 kbps for different cell loss distribution: uniform, binomial, and Poisson. These figures depict how the predicted MOS scores (obtained from the corresponding BSDB values) vary with cell loss rate when the lost cell was replaced by silent samples.

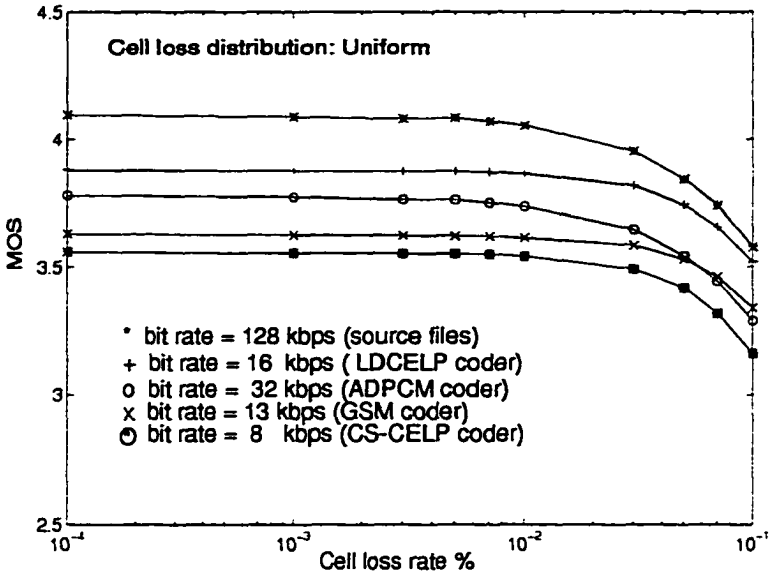


Figure 4.2.1.1 Predicted MOS versus cell loss: Uniform distribution.

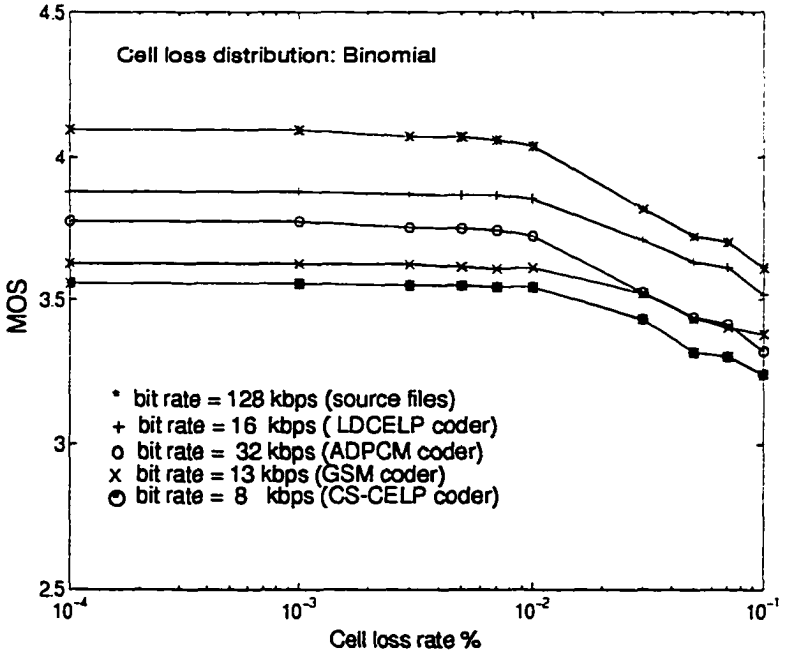


Figure 4.2.1.2 Predicted MOS versus cell loss: Binomial distribution.

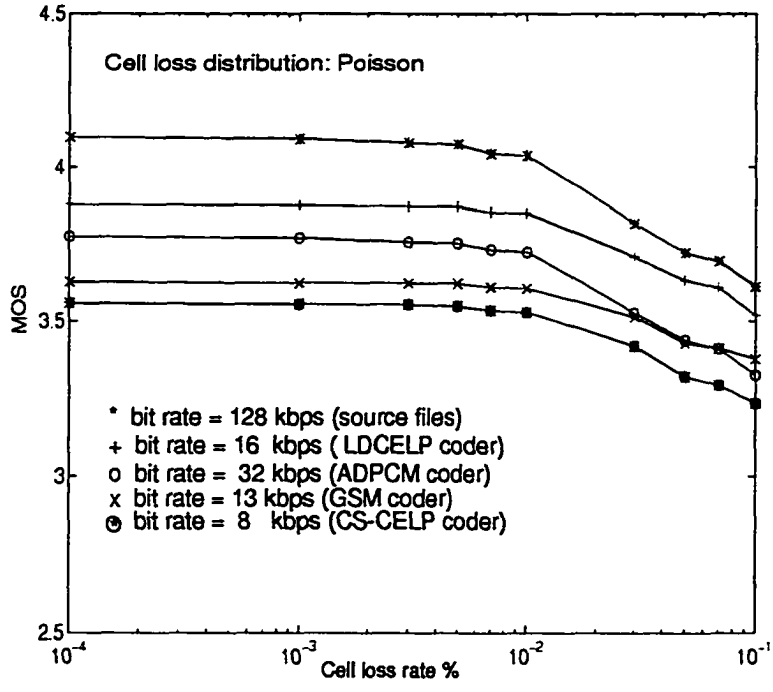


Figure 4.2.1.3 Predicted MOS versus cell loss: Poisson distribution.

Figures 4.2.1.1–4.2.1.3 show, as expected, that the speech quality decreases with increase of the cell loss rate. For example, speech quality at 10^{-4} (which we consider as zero cell loss) are 4.1, 3.77, 3.88, 3.63, and 3.56 for the source files, ADPCM coder, LDCELP coder, GSM coder, and CS-CELP coder respectively while the corresponding MOS values for 10% cell loss, at uniform cell loss distribution are 3.58, 3.29, 3.52, 3.34, and 3.16 respectively. Figures 4.2.1.1–4.2.1.3 show that with 10% cell loss rate, which is assumed to be a worst case in private ATM networks [43], the quality is kept above 3.2 for bit rate 8 kbps. To study the degradation behavior for each bit rate (coder algorithm), we calculate the degradation in speech quality, taking MOS at zero cell loss as a reference point, versus the cell loss as shown in Figs. 4.2.1.4–4.2.1.6 for the three cell loss distributions.

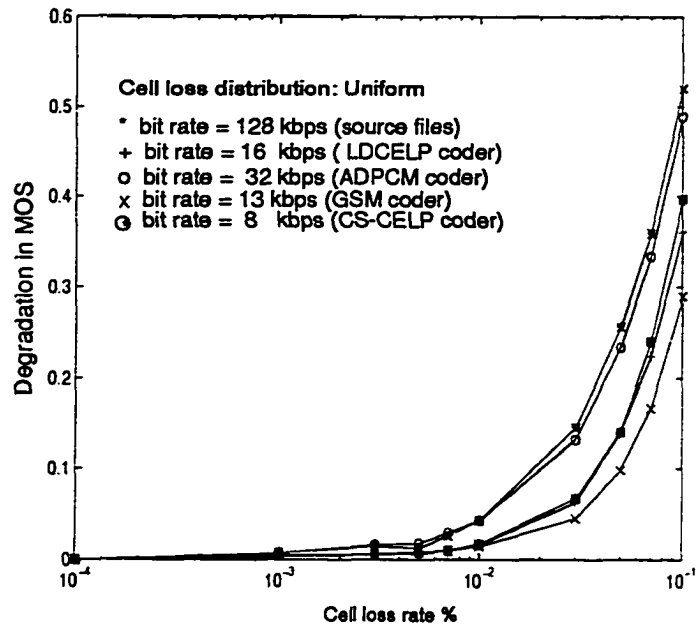


Figure 4.2.1.4 Degradation of speech quality versus cell loss: Uniform distribution.

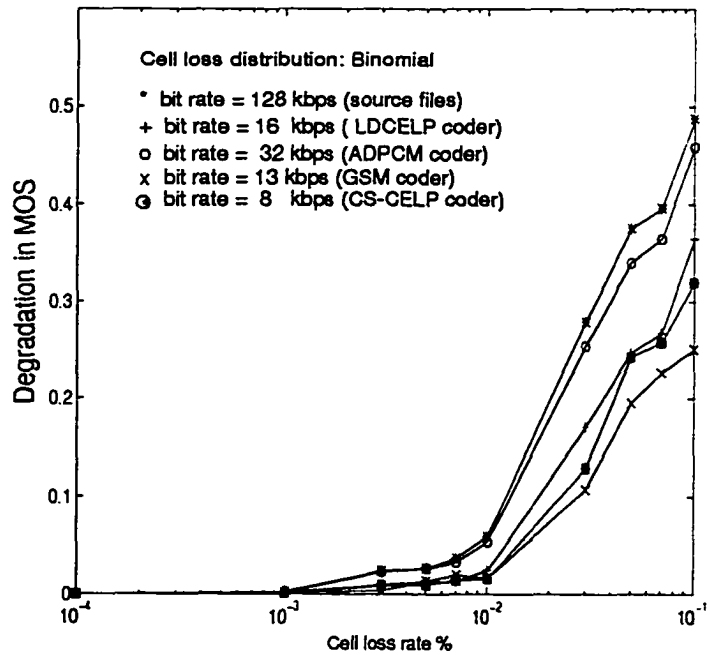


Figure 4.2.1.5 Degradation of speech quality versus cell loss: Binomial distribution.

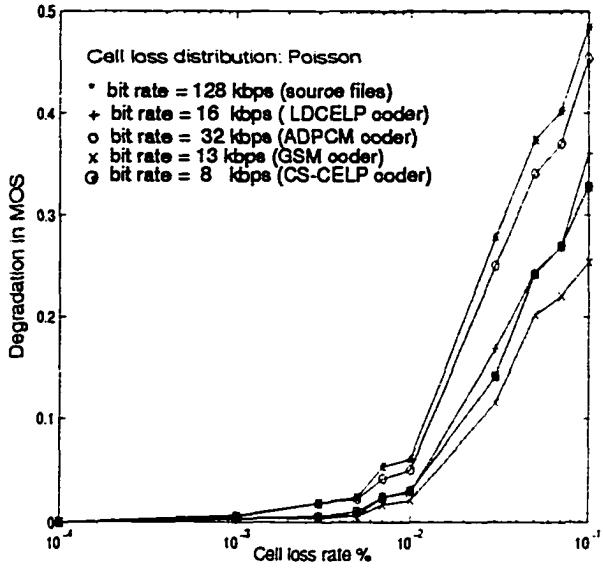


Figure 4.2.1.6 Degradation of speech quality versus cell loss: Poisson distribution

Figs. 4.2.1.4–4.2.1.6 shows that the degradation rate increases with the increase of coding bit rate for bit rate 128, 32, 16, 8 kbps and the lowest degradation rate is for GSM coder at 13 kbps. We believe that when the bit rate is high, the speech utterances are clear and the user can easily perceive the degradation effect, while for the lower bit rate, the speech utterances are not so clear so that the user can't easily distinguish the degradation effect for small variation in cell loss. A careful study of figures 4.2.1.1–4.2.1.6 shows that the degradation behavior of different bit rate coders are same for the three cell loss distribution assumptions (uniform, binomial, and Poisson). Figures 4.2.1.7–4.2.1.9 depict the variation of MOS versus the cell loss for certain bit rate coder under different loss distributions. Figures 4.2.1.7–4.2.1.9 illustrate that the speech quality doesn't strongly depend on the cell loss distribution, but it mainly depends on the value of the cell loss itself. Thus, we can normalize the results for the different cell loss distributions as described in Figs 4.2.1.10 and 4.2.1.11.

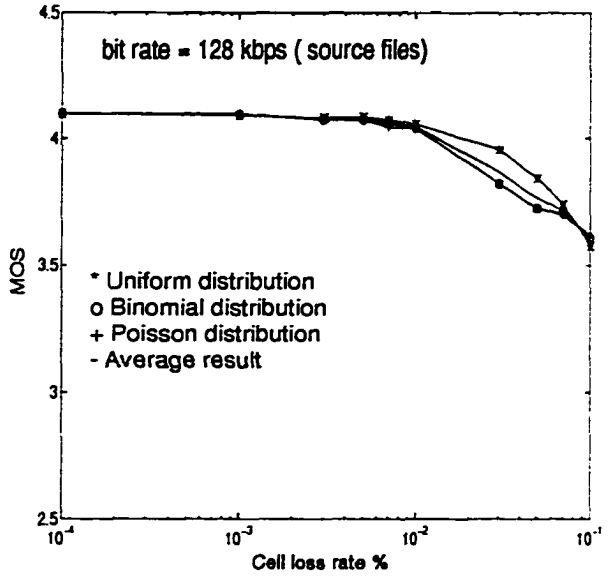


Figure 4.2.1.7 Degradation of speech quality of 128 kbps bit rate versus cell loss for different cell loss distributions

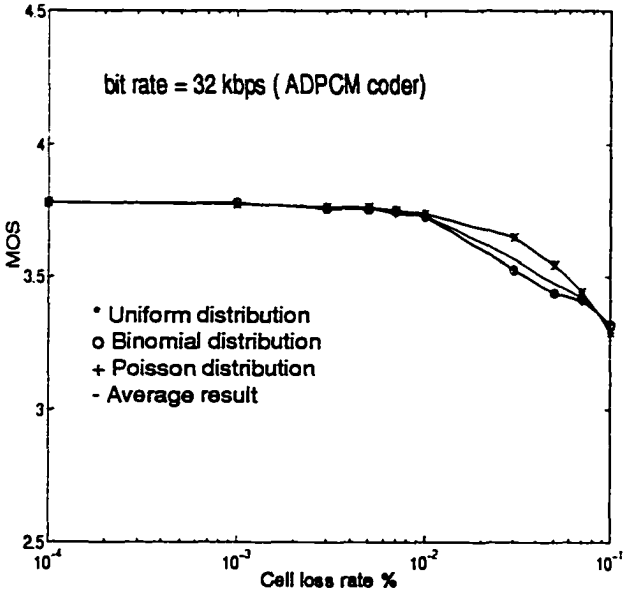


Figure 4.2.1.8 Degradation of speech quality of 32 kbps bit rate versus cell loss for different cell loss distributions.

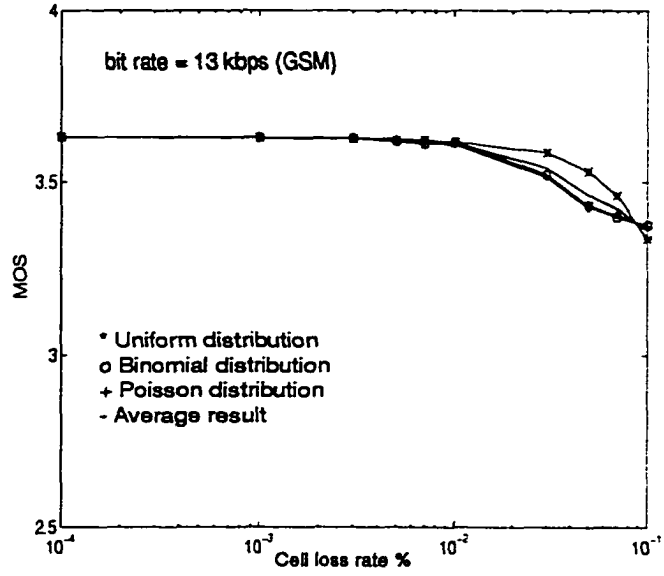


Figure 4.2.1.9 Degradation of speech quality of 13 kbps bit rate versus cell loss for different cell loss distributions.

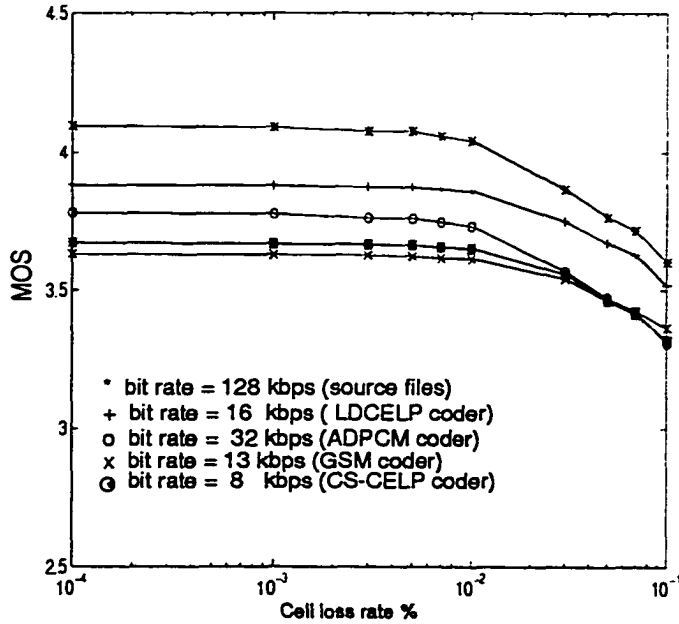


Figure 4.2.1.10 Average MOS values versus cell loss for different bit rate coding.

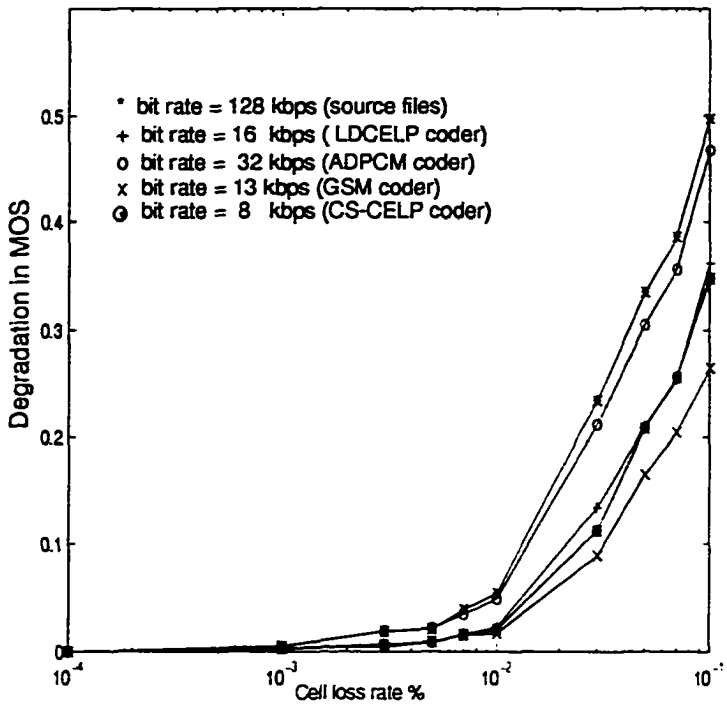


Figure 4.2.1.11 Average degradation in the speech quality versus cell loss for different bit rate coding.

The results from Figs 4.2.1.10 and 4.2.1.11 is in consistent with the previous results (Fig. 4.2.1.4–4.2.1.6) and can be used to study the degradation effect of cell loss on the speech quality for different coders.

4.2.1.2 Replacement of the lost cell by the previous successfully received one

For the second replacement technique, in which the lost cell is replaced by the previous successfully received one, we repeat the previous study (done in the first replacement technique) and we get the same behavior results as for the first replacement technique. But the second algorithm shows improvements in the speech quality with respect to the first one. For example, we choose a 32 kbps bit rate to compare the speech quality

differences in case of using the two replacement techniques. This comparison is depicted in Fig. 4.2.1.12.

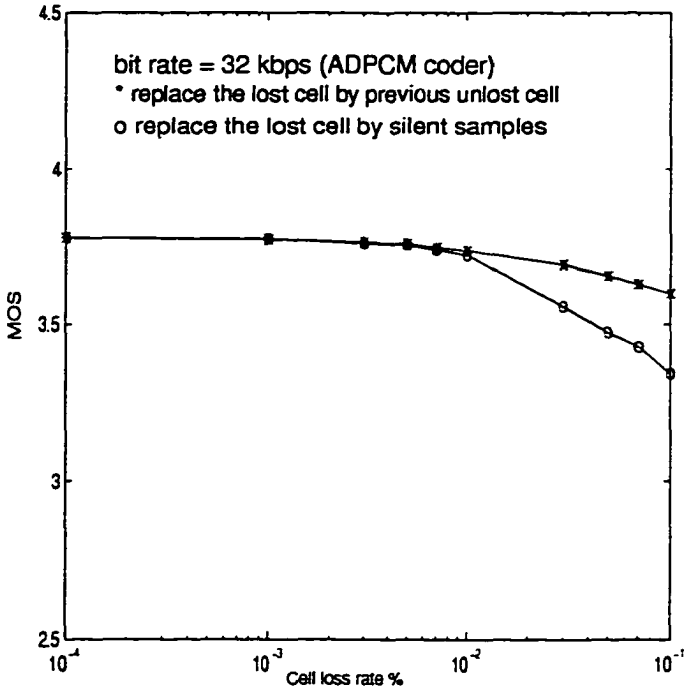


Figure 4.2.1.12 Average MOS for a 32 kbps bit rate for the two replacement techniques.

It is clear that for the same cell loss, second replacement technique gives a higher speech quality. For example at 10% cell loss, MOS value for the first replacement technique is 3.31 while it is 3.6 for the second replacement technique. Instead of introducing the improvement effect of the second replacement technique for every bit rate (coder algorithm), we plot the improvement effect of the second replacement technique, for all coders, with cell loss variation as shown in Fig. 4.2.1.13 . In summary, Fig. 4.2.1.13 demonstrates that the second replacement technique produce better results when compared with the first one.

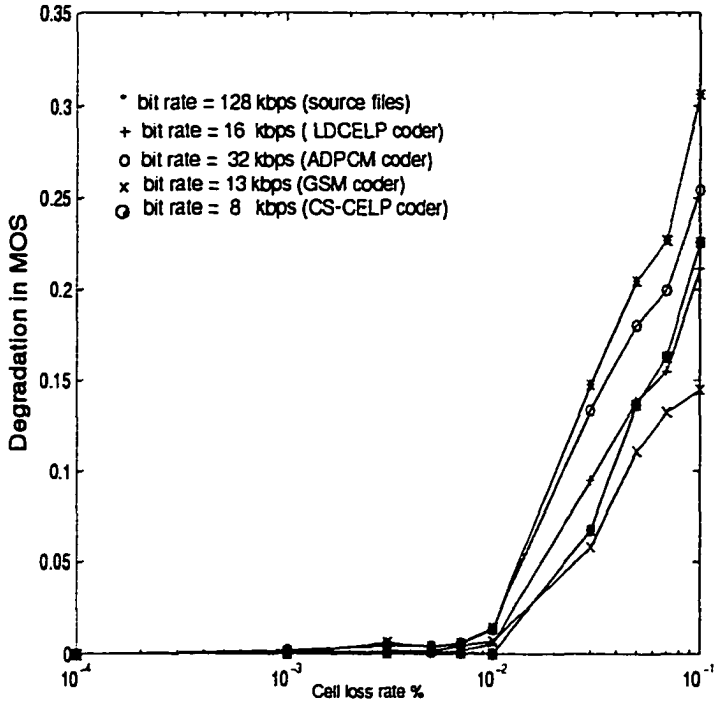


Figure 4.2.1.13 Improvement effect of using the second replacement technique rather than using the first one.

4.3 Conclusion

This chapter has presented a discussion of the issues involved in predicting the degradation impact of cell loss on speech quality over ATM network. From speech designing point of view, for given speech coding algorithms, our techniques can be used to predict the quality performance of speech coding algorithms due to cell loss impairments that will be introduced by ATM networks. Prediction the speech quality over ATM network help in designing the speech coders and controlling their electrical parameters to maintain certain speech quality.

From network point of view, the proposed technique can be used as a tool to predict the performance of speech reconstruction algorithms, that deal with the cell loss problem,

and choose the reliable methods. Also degradation information produced by the proposed technique can be used to aid in designing of the management, congestion control protocols and assignment rules that allows the meeting of connection performance standards and the achieving of certain quality of service (QOS) requirements. The study also shows that up to 10% of speech cells can be lost while keeping the speech quality over MOS (Mean Opinion Score) of 3.2 for some coders such as LDCELP at 16 kbps, ADPCM at 32 kbps, GSM at 13 kbps, and CS-CELP at 8 kbps. In summary, the proposed technique allows to predict the performance behavior of the suggested coding algorithms to be used over ATM networks.

Chapter 5 IMPACT OF JITTER ON SPEECH QUALITY OVER ATM NETWORK

5.1 Overview

In the packet-switched network (e.g., Asynchronous Transfer Mode, ATM), traffic form all sources is packetized, and statistical multiplexing techniques are used to combine all network traffic through a single switching fabric. This allows higher network utilization but requires more sophisticated controls to ensure that the appropriate QOS is provided [48]. In ATM networks, the most important QOS parameters are those dealing with cell loss, cell delay, and cell delay variation (delay jitter) [49]. Constant bit rate (CBR) traffic sources, e.g., CBR audio and video, will be supported by the ATM network, and in fact it is expected they will comprise a major portion of traffic on the network [50]. One of the central problems in the support of real-time applications (voice, video) within ATM networks is the existence of delay jitter [51]. The delay time of a packet in ATM networks is composed of a fixed component of propagation delay and a variable component caused by the waiting time in the buffers of the network [51]. Delay jitter is the variation of the delays with which packets travelling on a network connection reach their destination [52]. For good quality of reception, continuous-media streams require that the jitter be kept below a sufficiently small upper bound [52]. Thus, the network's contribution to jitter should be as small as is economically feasible [53] and upper bounds on cell loss and jitter are needed to ensure any desired level of output quality. The previous study [54], explained in the previous chapter, shows that up to 10% of speech packets can be lost over ATM networks while keeping the speech qualify over MOS (Mean Opinion Score) of 3.2 for some speech coders. Our objective in this chapter is to predict the

user's opinion of speech quality due to jitter introduced in ATM networks. The user's opinion is important for the proper design of network algorithms such as routing, flow control, and management techniques. In the emerging technology, the user expects a minimum guaranteed value of QOS regardless of traffic intensity, service variety, or network imperfections [55]. A careful definition of the user requirements would also greatly assist in the design of future telecommunication systems, services [1], and audio applications [56].

Since introducing jitter for the speech signals will change their time characteristics, we can't use the previous objective techniques [57] and [58], introduced in chapters 2 and 3, in predicting the ATM network's output speech quality because the time matching between the input and the output signals is necessary in that technique.

To study the impact of jitter on speech quality, we have introduced a new perceptually-output-based objective technique to predict the output speech quality without referring to the input speech. In this technique, we emulate several known features of perceptual processing of speech sounds by human ear (including critical-band masking, equal loudness, and the intensity-loudness power law operations) to map the speech power spectrum, P(f) in the frequency domain , into auditory power spectrum, B(z) in the bark domain, which is used to derive the perceptual LPC coefficients. Then, we use the radial basis functions neural network to map the perceptual LPC into the corresponding estimated speech quality.

The rest of this chapter is organized as follows: section 2 summarizes the calculation of the perceptual LPC parameters that are used to predict the output speech quality. The proposed evaluation system is illustrated in section three. Section four illustrates the

validity of the proposed objective technique. In section five, we study the impact of jitter on speech quality while conclusion and further work are introduced in section six.

5.2 Calculation of the Perceptual LPC Parameters

The speech frame is weighted by Hamming window and the consecutive frames overlapped by 50 %. By computing the magnitude square FFT spectrum, the frame's power spectrum P(f) is calculated and followed by some speech processing operations similar to what we did in chapter 2. These operations are summarized as follows:

- Compute the FFT-based sample spectrum,
- Warp the frequency axis into a Bark frequency scale,
- Apply a critical-band masking curve, i.e., convolve the Bark spectrum with a critical-band filter and sample at 1-Bark intervals,
- Weight the samples for equal loudness to adjust for the ear's frequency dependent loudness sensitivity,
- Apply 1/3 power law to simulate the non-linear relation between sound intensity and perceived loudness.

After performing the above operations, we obtain thirteen samples of the auditory spectrum, B(i), that covers the telephony spectrum (300–3400 Hz).

- The inverse DFT is applied to B(i) to yield the autocorrelation function, r_{ij} (r_j), dual to B(i).
- The perceptual LPC coefficients, α_i, are obtained by solving the Yule-Waker equation, $\sum_{i=1}^{p} \alpha_i r_{ij} = r_j$, with p (=10) is the order of LPC analysis.

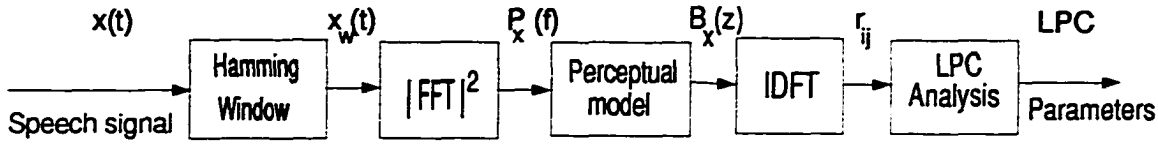


Figure 5.2.1 Basic transformations used in obtaining the perecptual LPC parameters.

5.3 Evaluation System

Like the technique illustrated in chapter 3, we prepare two evaluation systems, each for one sentence, that use radial basis function neural networks. Unlike the previous technique, this technique use LPC parameters of the speech frames of the speech signal, instead of the cepestrum distance between the input and the output frames, in predicting the speech quality.

To have a constant feeding input to the neural network, we divide each sentence into fixed number of frames. Typically we use 211 overlapping frames with their sizes in the range of 20 to 30 msec. We believe that dividing the sentence into fixed number of frames, with different size, and then approximate that frame by all-pole model, with a suitable order, is more accurate than using the time warping technique that modifies the time shape of the speech signal. The output speech sentence (S1 and S2) that processed by 10 different coders are used to prepare the learning data for the two sub-evaluation systems.

During the learning phase of our evaluation systems, the actual output quality scores is compared to the desired scores and the errors between the actual and desired scores are then used to determine the best network coefficients that minimize the predicted square error (PSE) for certain value of spread constant, sc. The learning procedures is repeated for different values of sc. Figure 5.3.2 shows the PSE during the learning phase of one the sub-evaluation system while Fig.5.3.3 depicts its production phase.

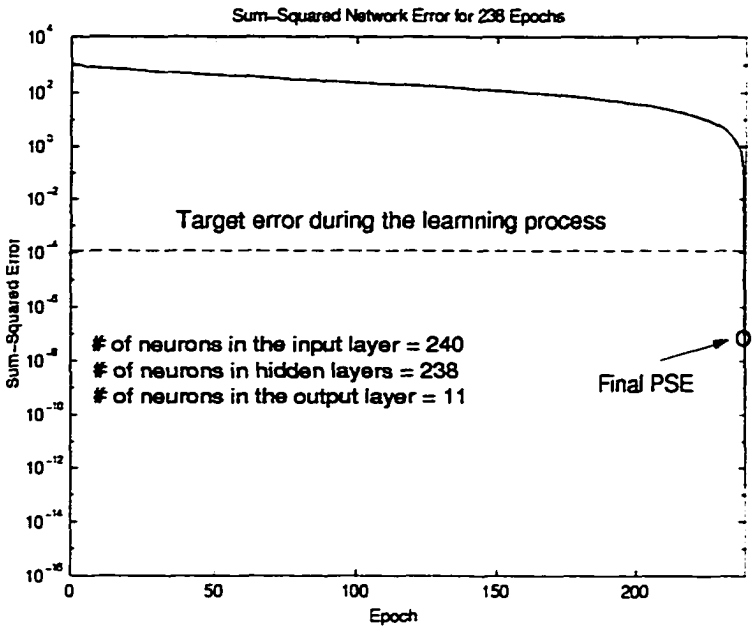


Figure 5.3.2 Predicted square error for the learning phase of the radial basis functions neural network.

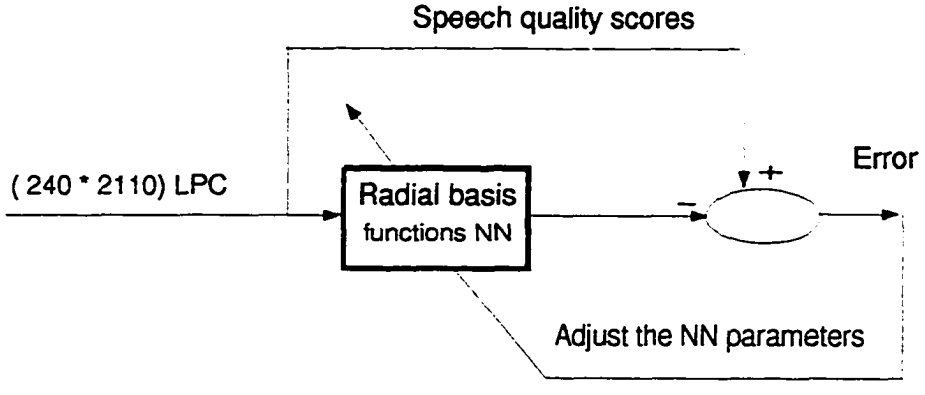


Figure 5.3.3 Learning phase for one of the sub-evaluation system.

At the end of the learning phase, we have some evaluation system, each achieved mapping of PCDF into the predicted speech quality scores. We tested all of these systems to predict the average MOS of other 16 coders. We chose the one that gave the most accurate result for the two sub-evaluation systems. So our evaluation system consists

of the best subvaluation systems. This evaluation system can be used to predict the performance of speech coder over ATM network, voice communication systems, and mobile communication systems.

5.4 Validity of the Proposed Evaluation System

The validity of the proposed evaluation systems is checked by comparing the average predicted MOS obtained from our technique to those obtained from the subjective test. The actual MOS and the predicted one, based on the mixed speakers, is illustrated in Fig. 5.4.4 and Fig. 5.4.5 for the two neural networks (sub-evaluation systems) respectively. Figure 5.4.6 compares the total predicted MOS, the average score of the two networks, with the actual MOS values.

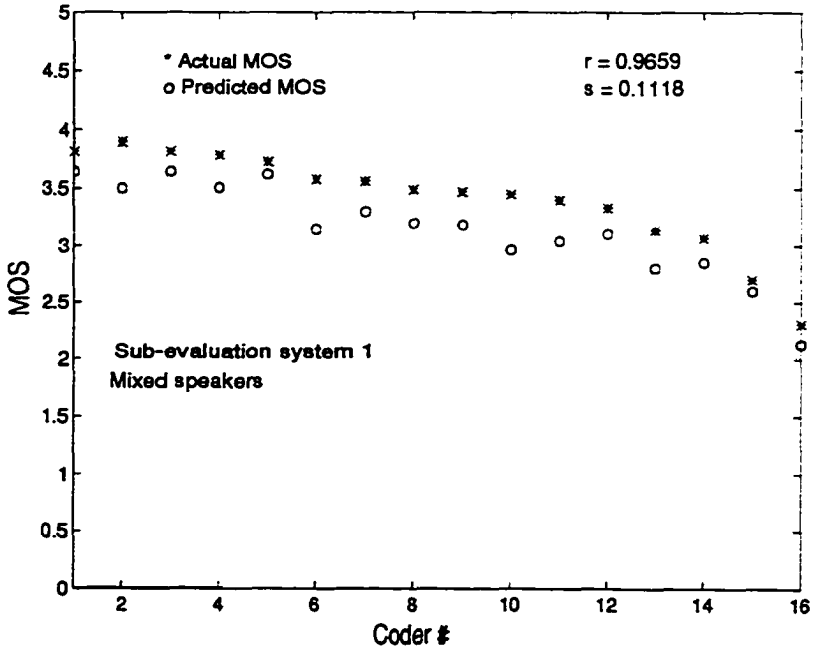


Figure 5.4.4 Actual and predicted MOS values for sub-evaluation system 1.

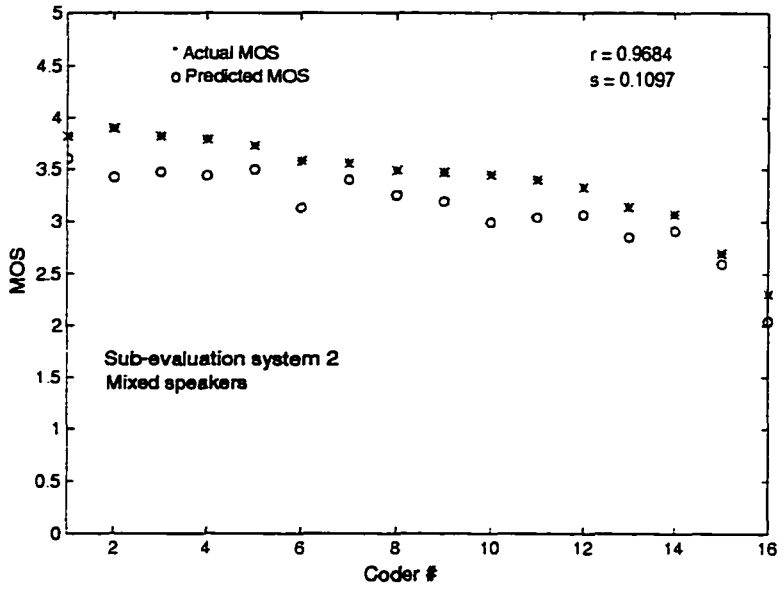


Figure 5.4.5 Actual and predicted MOS values for sub-evaluation system 2.

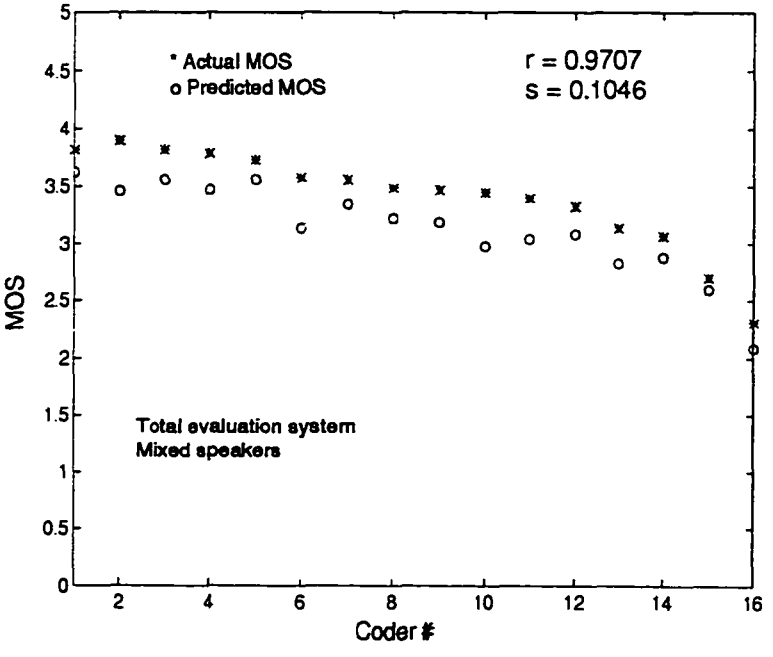


Figure 5.4.6 Actual and predicted MOS values for the total evaluation system.

The results of Figs. 5.4.4–5.4.6 show that our objective technique is reliable in predicting the speech quality and it is highly correlated with the subjective test. We believe this technique is suitable in studying the impact of jitter on the speech quality, especially when the speech materials, that used in training the neural networks, are used in that study.

5.5 Impact of Jitter on Speech Quality

As the packets (cells) pass through the ATM network, each can encounter a varying amount of queueing delay in the statistically multiplexed links. The delay experienced by a packet can be divided into a fixed delay, D, which is the same for each packet in a call, and a variable delay, W. The fixed delay arises from the propagation of packets, processing time at sender and receiver, and from a fixed buffering delays in the network. The variable delay results from queuing and from other variable processing delays in the network [59]. The alternation of the initial periodic nature of a constant bit rate cell stream due to such delays is the phenomenon of jitter [60]. The impact of jitter in the output speech can be simulated by inserting silent samples between the ATM cells. For example inserting one silent sample between two consecutive cells is equivalent to introduce a time jitter of 0.125 ms (i.e. , sampling time) between those two cells. Thus, at the receiver end, the listener has the degraded speech and the goal of this chapter [61] is predicting the speech quality as perceived by the user. In this study, we choose three different distributions for the jitter: uniform, binomial, and Poisson. For each distribution, the mean values of jitter are 0.5, 1.5, 2.5, 3.5, and 5 ms. Figures 5.5.7–5.5.9 shows the jitter distributions at 5 ms jitter mean value.

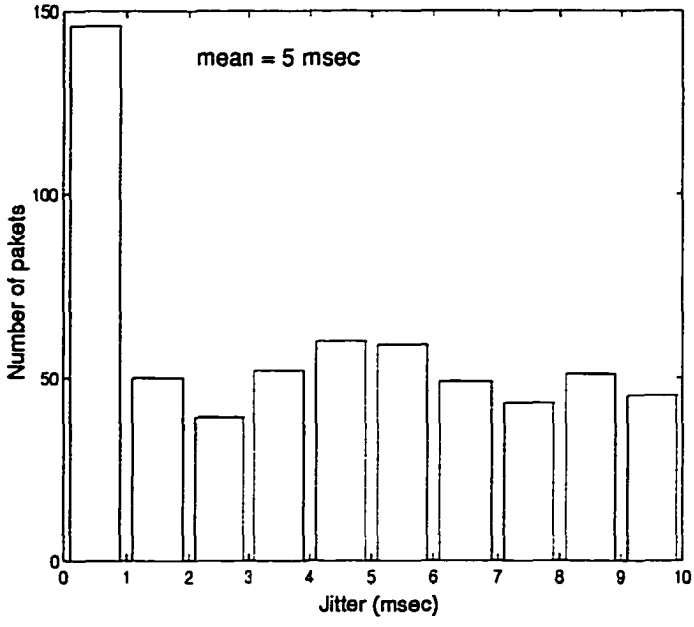


Figure 5.5.7 Uniform jitter distribution.

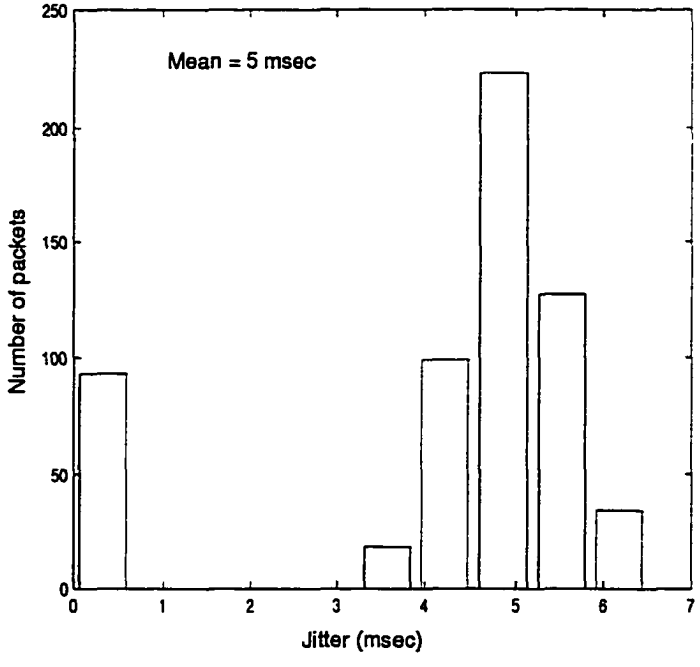


Figure 5.5.8 Uniform jitter distribution.

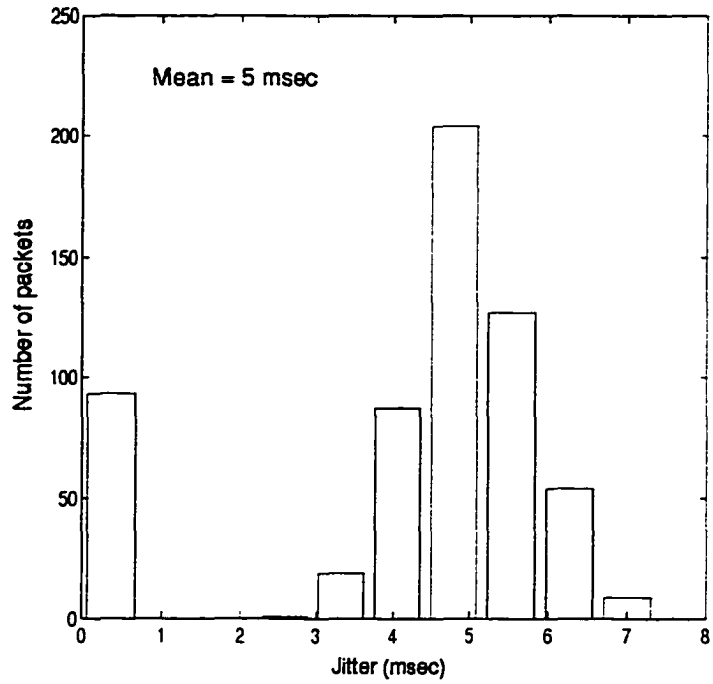


Figure 5.5.9 Poisson jitter distribution.

For certain bit rate and a given coder algorithm, we apply the same jitter effect parameters (distribution, mean value, number of cells that faced the jitter) on the 24 speech files. Then we calculate the perceptual LPC coefficients of the output speech files that used to feed the trained radial basis functions neural network. The neural network's output represents the predicted average MOS score of the jitter effect for the chosen bit rate coding algorithm.

5.5.1 Numerical results

In this section, impact of jitter on speech quality in ATM networks is analyzed. Figures 5.5.1.10–5.5.1.12 show the relation between degradation in speech quality versus jitter for the three distributions.

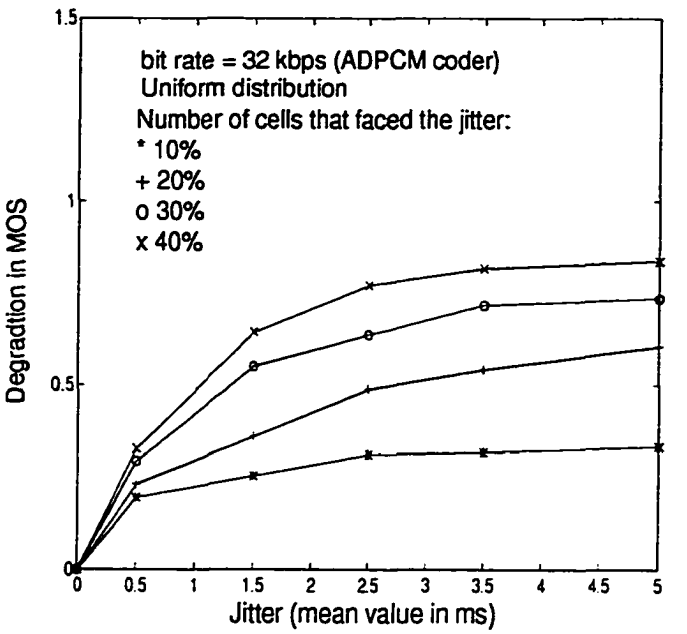


Figure 5.5.1.10 Degradation of speech quality versus jitter: Uniform distribution.

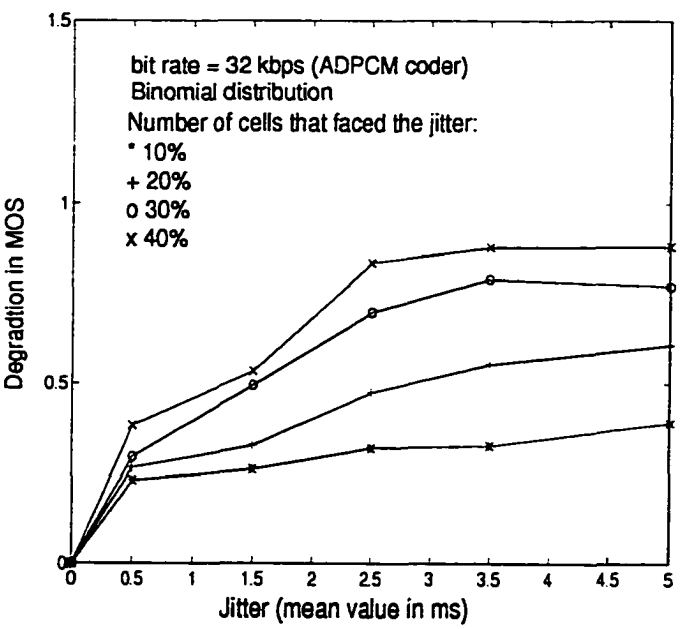


Figure 5.5.1.11 Degradation of speech quality versus jitter: Binomial distribution.

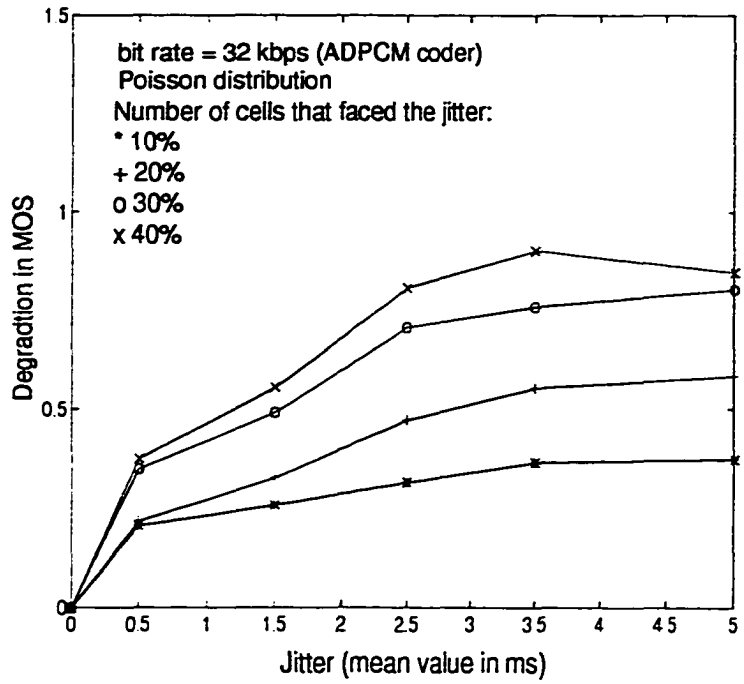


Figure 5.5.1.12 Degradation of speech quality versus jitter: Poisson distribution.

Results of Figs. 5.5.1.10–5.5.1.12 show that the degradation in speech quality, as expected, increases with increasing the jitter and the degradation behavior is almost same for different distribution. This indicates that the degradation depends most likely on the value of jitter itself and its variation doesn't strongly depend on the jitter distribution. This idea is illustrated in Figs. 5.5.1.13–5.5.1.16.

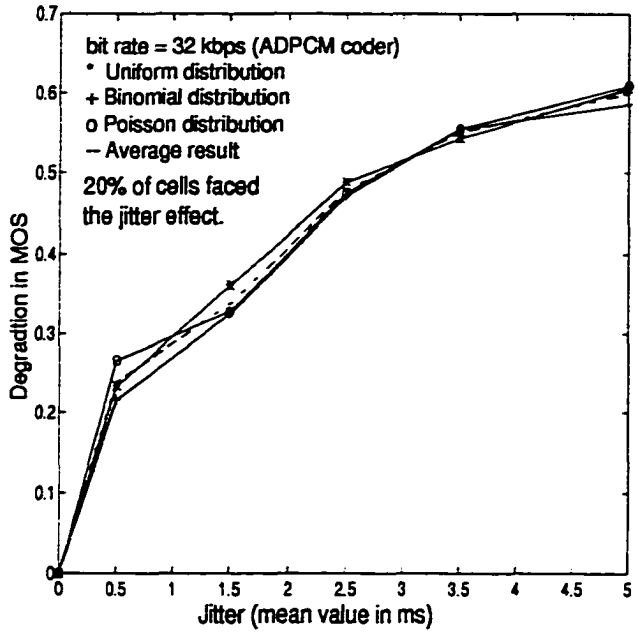


Figure 5.5.1.13 Variation of speech quality versus jitter for bit rate = 32 kbps and 20% of cells face jitter.

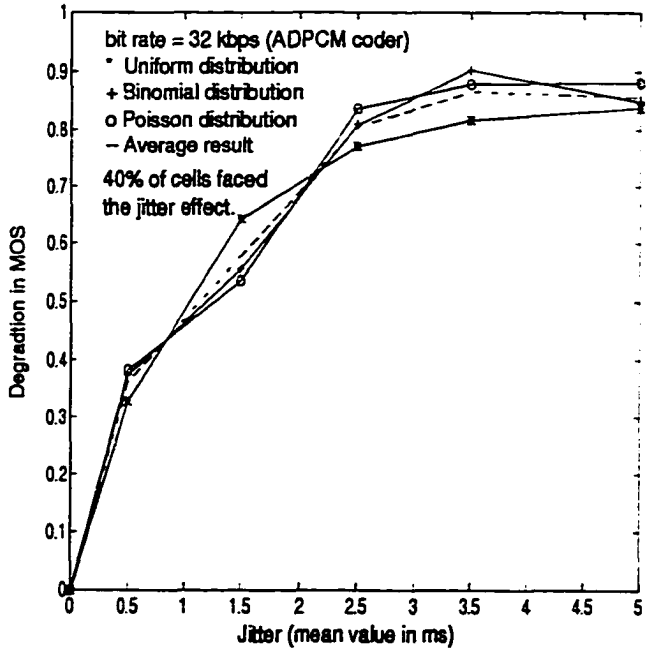


Figure 5.5.1.14 Variation of speech quality versus jitter for bit rate = 32 kbps and 40% of cells face jitter.

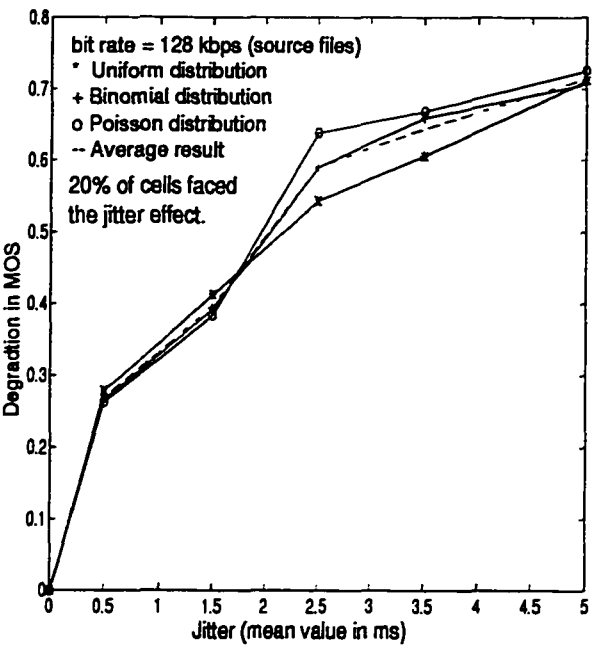


Figure 5.5.1.15 Variation of speech quality versus jitter for bit rate = 128 kbps and 20% of cells face jitter.

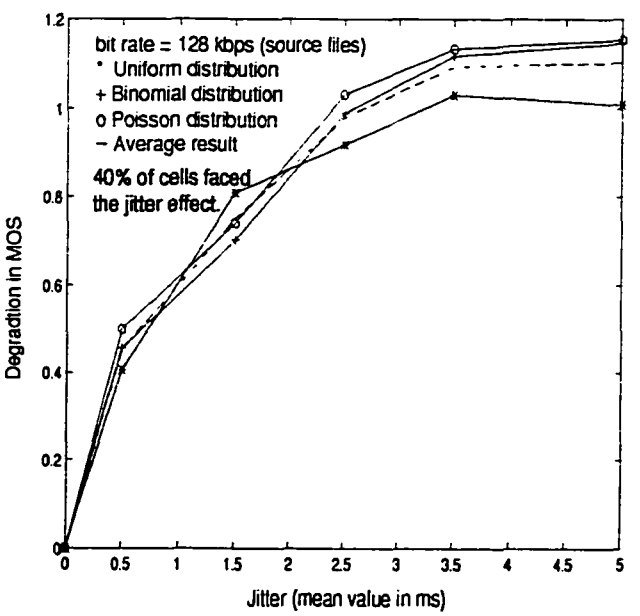


Figure 5.5.1.16 Variation of speech quality versus jitter for bit rate = 128 kbps and 40% of cells face jitter.

Figs. 5.5.1.13–5.5.1.16 shows that the degradation variation due to jitter distribution is within 0.1 and this indicates that normalization of the result, with respect to jitter distribution, will not introduce a high prediction error. In the following figures, we normalize the result with respect to the jitter distribution. Figures 5.5.1.17–5.5.1.21 depict how the predicted MOS varies with jitter value (independent of jitter distribution) for different bit rate coding.

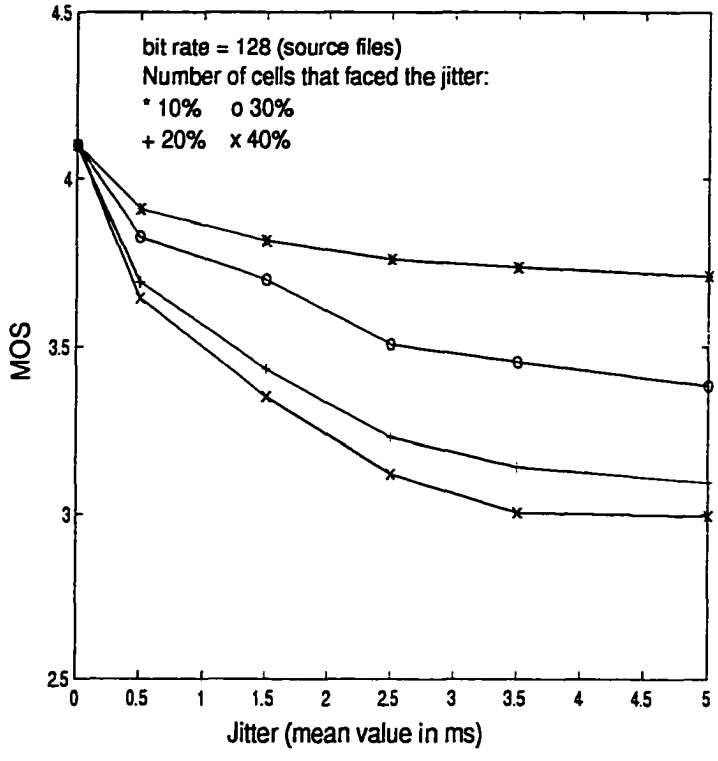


Figure 5.5.1.17 Variation of speech quality versus jitter for 128 kbps bit rate.

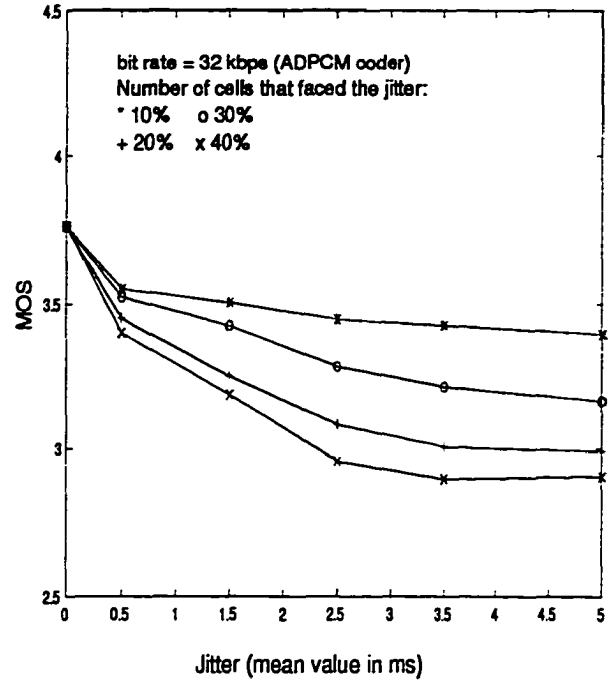


Figure 5.5.1.18 Variation of speech quality versus jitter for 32 kbps bit rate.

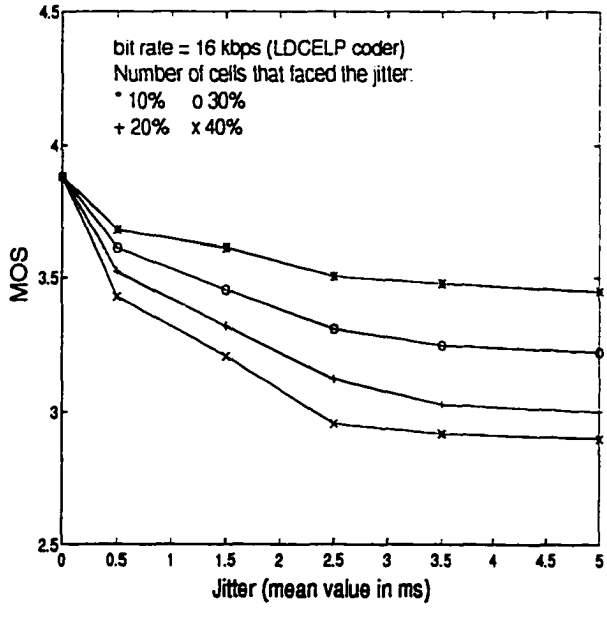


Figure 5.5.1.19 Variation of speech quality versus jitter for 16 kbps bit rate.

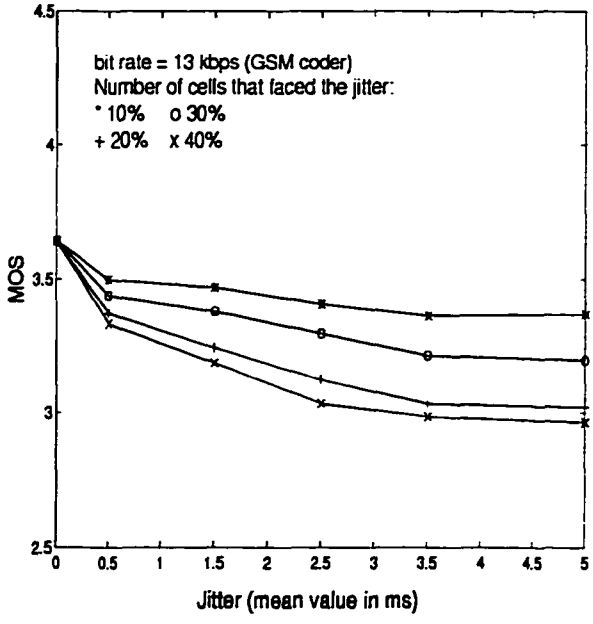


Figure 5.5.1.20 Variation of speech quality versus jitter for 13 kbps bit rate.

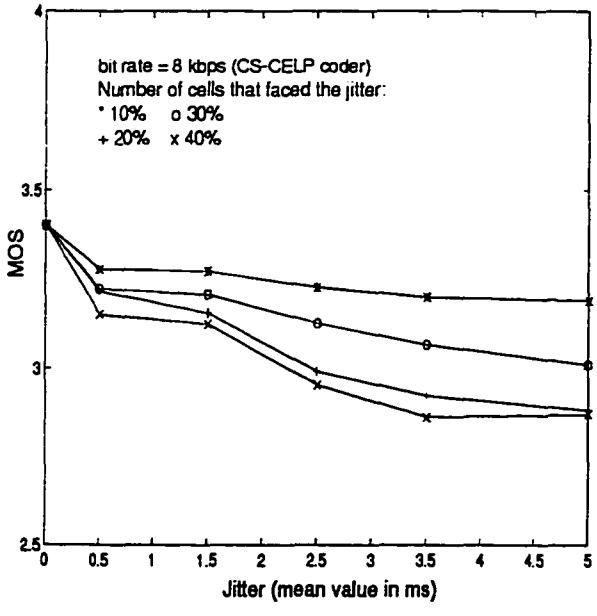


Figure 5.5.1.21 Variation of speech quality versus jitter for 8 kbps bit rate.

It is clear that, we can use Figs. 5.5.1.17–5.5.1.21 in obtaining the jitter tolerance value , for each bit rate coder, to achieve certain speech quality (MOS). For example to keep MOS at 3.2 for bit rate 128 kbps, the jitter introduced from ATM networks should be 2.5 msec and number of cells (cell ratio) which will face that jitter can be up to 30%. If cell ratio equal to 10% or 20%, the jitter limit can be more than 5 msec.

The above study shows that, by using the proposed technique, we can predict the quality performance of speech coding algorithm due to jitter impairments that will be introduced by the ATM networks. Furthermore, the proposed technique can be used to assign upper jitter limit for the suggested coding algorithms to be used over ATM networks. Also, our technique can be used to select among jitter control schemes that absorbs the jitter by buffering data at the destination.

5.6 Conclusion

This chapter has presented a discussion of the issues involved in predicting the degradation impact of jitter on speech quality over ATM networks. From speech coder designing point of view, for given speech coding algorithms, the proposed technique can be used to predict the quality performance of speech coding algorithms due to jitter impairments that will be introduced by ATM networks. Prediction the speech quality over ATM networks helps in designing the speech coders and controlling their electrical parameters to maintain certain speech quality.

From network design of view, the proposed technique can be used to assign upper jitter limit for the suggested coding algorithms to be used over ATM networks. Furthermore, it can be used to select among jitter control schemes that absorbs the jitter by buffering data at the destination. Also degradation information produced by the proposed technique

can be used to aid in designing of the management, congestion control protocols and assignment rules that allowed the meeting of connection performance standards and the achieving of certain quality of service (QOS) requirements.

Although we have studied only the impact of jitter on speech quality, major research is still necessary in this area to predict the degradation impact of both the cell loss and jitter, in the same time, on speech quality over ATM networks.

Chapter 6 CONCLUSION AND FURTHER WORK

Throughout this work, the problem of predicting speech quality using the objective measure techniques, was addressed. We have provided two preceptually-oriented objective measure techniques that are highly correlated to human responses across a wide range of quality levels and for a wide range of speech coding techniques. These new objective measure techniques emulate several known features of perceptual processing of speech sounds by human ear (including critical-band masking, equal loudness, and the intensity-loudness power law operations) to map the speech power spectrum into auditory power spectrum (bark domain). In the first objective measure technique, as illustrated in the second chapter, we use the auditory power spectrum in calculating the bark spectral distance per band (BSDB) between the input and the output coded speech signals. Then, we use the abductive networks, that evolved from neural network, statistical modeling, and artificial intelligence concepts, to estimate the speech quality from the BSDB. In the third chapter, we introduced the second objective measure technique in which we derived the perceptual LPC coefficients from the auditory spectrum. The perceptual LPC coefficients are used to calculate, for each frame, the cepstrum distance between the input and the output coded speech signals. Then, we use the radial basis functions neural network to map the perceptual cepstrum distance per frame into the corresponding estimated speech quality.

After extensive experimentations and validation of our techniques, we obtained results of r and s in the range of 0.96 and 0.1 respectively where r is the correlation coefficient between subjective test ratings and corresponding objective measures, and s is the standard deviation of the prediction error. The results indicate that our proposed techniques are

reliable and robust in evaluating the coded speech quality.

Our objective measure techniques can be used in development, testing, refinement, deployment, or standardization of algorithms and equipment that process speech signals. Such techniques help to minimize the number of costly and time consuming formal subjective tests required by those activities including the performance evaluation of the speech coders.

Furthermore, in this research, we have presented a discussion of the issues involved in predicting the user's opinion of speech quality due to cell loss and jitter introduced in ATM networks. The user's opinion is important for the proper design of network algorithms such as routing, flow control, and management techniques.

We have used the first objective technique in studying the impact of cell loss on speech quality over ATM networks. Moreover, we compare the results between two different cell loss's replacement techniques: stuffing silent samples and inserting the previous information in the lost cell. Study shows that the second replacement techniques produces better result when compared with the first one. The study also shows that up to 10% of speech cells can be lost over ATM networks while keeping the speech quality over MOS (Mean Opinion Score) of 3.2 for some speech coders.

Since introducing jitter for the speech signals will change their time characteristics, we can't use the previous objective techniques in predicting the ATM network's output speech quality because the time matching between the input and the output signals is necessary in that technique. To study the impact of jitter on speech quality, we have introduced a new perceptually-output-based objective technique to predict the output speech quality without referring to the input speech. In this technique, we have used the auditory power spectrum

in deriving the perceptual LPC parameters. Then, we use the radial basis functions neural network to map the perceptual LPC into the corresponding estimated speech quality.

From speech coder designing point of view, prediction the speech quality over ATM networks helps in designing the speech coders and controlling their electrical parameters to maintain certain speech quality.

From network design point of view, the proposed techniques can be used to assign upper cell loss and jitter limits for the suggested coding algorithms to be used over ATM networks. Also degradation information produced by the proposed techniques can be used to aid in designing of the management, congestion control protocols and assignment rules that allowed the meeting of connection performance standards and the achieving of certain quality of service (QOS) requirements.

Although we have mainly studied the impact of cell loss and jitter on speech quality in two separate ways but our studying help, to some point, in understanding the impact of those two problems (cell loss and jitter) on the speech quality. Major research is still necessary in this area to predict the degradation impact of cell loss, jitter, end to end delay, and echo on speech quality over ATM network in a real time environment.

Appendix A Linear Predictive Coding (LPC) Analysis

Linear predictive coding (LPC) can provide a complete description for a speech production model. The basic idea underlying LPC is that each discrete speech sample, x_t, can be represented as a linear combination of previous samples, and prediction errors can then be minimized according to the mean-square value of the prediction error, e_t, which is defined by [9]:

$$e_t = x_t + \sum_{i=1}^{p} \alpha_i \, x_{t-i} \tag{A.1}$$

where p is the order of LPC analysis, and α_i are LPC coefficients. The LPC coefficients which minimize the mean square prediction error can be obtained by setting the partial derivative of the mean-square prediction error (with respect to each α_i) equal to zero. Apply the z-transform to eq. (A.1), the following expression is obtained:

$$E(z) = \sum_{i=0}^{p} \alpha_i \; z^{-i} \, X(z) \tag{A.2}$$

Let us denote $H(z)$ as follows:

$$H(z) = \frac{1}{\sum_{i=0}^{p} \alpha_i \, z^{-i}} \tag{A.3}$$

Then eq. (A.2) is expressed as:

$$X(z) = H(z) \, E(z) \tag{A.4}$$

The spectra of e_t and x_t are obtained by setting $z = e^{j w T}$ where T is the sampling time. Since the denominator of $H(z)$ has p complex roots, the $H\left(e^{j w T}\right)$ has $p/2$

resonant frequencies, which correspond to formant frequencies. This implies that the LPC technique can model the spectrum of the vocal tract as a spectrum of an order-p model $H(z)$. The expression of eq. (A.4) indicates that the spectrum $X(z)$ of discrete speech samples is produced as the product of the spectrum $H(z)$ of the vocal tract and the spectrum $E(z)$, which is the spectrum of the unpredictable signal formed from the past p speech samples. Therefore the $E(z)$ corresponds to the spectrum of voice excitation. Suppose that the voice excitation is white noise in the case of unvoiced speech, and an impulse in the case of voiced speech, as shown in Fig. A.1 so that the spectrum of voice

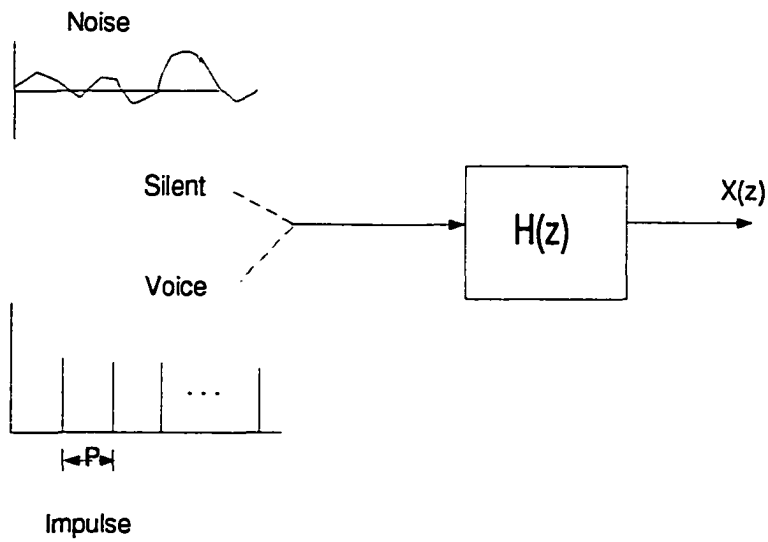


Figure A.1 Voice production model.

excitation $E(z)$ becomes spectrally flat. Under such conditions, the speech spectrum $X(z)$equals the spectrum at the vocal tract, so that $X(z) = G\,H(z)$ where G is the gain constant.

Appendix B Cepstral Analysis

As shown in appendix A, the basic model of speech production can be considered as a vocal tract filter $H(z)$ excited by a periodic excitation function $E(z)$ for voiced speech or white noise $E(z)$ in the case of unvoiced speech.

If, in the frequency domain, the product of the excitation and filter spectrum is transformed to the summation of these two spectra (logarithm operation), the transformation from the frequency domain back to the time domain by Fourier transform results in the *cepstrum*, which can represent the excitation and vocal tract separately. The parameter for *cepstrum* is called *quefrency* and is effectively a (pseudo) time domain parameter. The excitation locates at high *quefrency* owing to its periodic high frequency, and the vocal tract locates at low *quefrency* owing to its smoothed spectral envelope. This separable representation is very suitable to the deconvolution of speech and this analysis is called cepstral analysis [62], [63]. There are two types of cepstral analysis: FFT cepstral and LPC cepstral analysis [64], [65]. In the FFT cepstral analysis, a fast Fourier transform is directly applied to the speech signal. On the other hand, in the LPC cepstral analysis, the z-transform is applied to the speech signal modelled by LPC analysis. To investigate properties of the LPC *cepstrum*, the excitation $E(z)$ and the vocal tract filter $H(z)$, in the speech spectrum $X(z)$, are linearly separated by a complex logarithm operation applied to eq. (A.4). Then

$$log\, X(z) = log\, H(z) + log\, E(z) \tag{B.1}$$

The LPC cepstral coefficients c_n are defined as the inverse z-transform of the above log-spectrum $log\ X(z)$. This indicates that the characteristics of vocal tract and excitation are well represented separately in the cepstral coefficients. The higher order coefficients take the excitation property and the lower order coefficients take the vocal tract property. The cepstral coefficients, c_n, of the spectra obtained from LPC analysis can be computed recursively from the LPC coefficients, α_i, by [9]:

$$c_n = -\alpha_n - \sum_{i=1}^{n-1} \frac{n-i}{n} \alpha_i\, c_{n-i} \quad , \ n \geq 1 \tag{B.2}$$

where $\alpha_i = 0$ when $i > p$ (p is the order of LPC analysis).

Cepstral coefficients have been widely used in speech recognition [9]. A variety of speech recognition systems using cepstral analysis have been reported [66]. A distinctive advantage of the cepstral analysis is that correlation between coefficients is extremely small so that simplified modelling assumptions can be applied [9].

REFERENCES

[1] S. Wolf, C. A. Dvorak, R. F. Kubichek, C. R. South, R. A. Schaphost, S. D. Voran. "How will we rate Telecommunications system performance?". *IEEE Commun. Mag. vol. 29, pp. 23-29, Oct. 1991.*

[2] ITU-T Recommendation. "Subjective Performance Assessment of Telephone-Band and Wideband Digital Coders". *pp. 83, 1993.*

[3] S. Voran. "An Overview of ITS Research on Perception-Based Audio Quality Assessment". *report, NTIA/ITS, Boulder, Colorado.*

[4] M. R. Schroeder, B. S. Atal, and J. L. Hall. *"Objective measure of certain speech signal degradations based on masking properities of human auditory perception". Frontiers of Speech Communication. New York: Academic, 1979.*

[5] Y. Be'ery, Z. Shpiro, T. Simchony, L. Shatz, and J. Piasetzky. "An efficient variable-bit-rate low-delay CELP". *Advances in Speech Coding, B. S. Atal et al. , Eds. New York: Kluwer, 1990.*

[6] S. Wang, A. Gersho. "An objective measure for predicting subjective quality of speech coders". *IEEE J. Select. Areas Commun. vol. SAC-10, pp. 819-829, 1992.*

[7] N. Kitawaki, H. Nagabuchi, and K. Itoh. "Objective quality evaluation for low-bit-rate speech coding systems". *IEEE J. Select. Areas Commun. , vol. SAC-6, pp. 242-248, Feb. 1988.*

[8] D. Goodman and R. Nash. "Subjective quality of the same speech transmission conditions in seven different countries". *IEEE Trans. COmmun. , vol. COM-30, pp. 642-654, Apr. 1982.*

[9] S. Quackenbush, T. Barnwell, and M. Clements. "Objective Measures of Speech Quality". *Englewood Cliffs, NJ: Prentice Hall, 1988.*

[10] D. Klatt. "Prediction of perceived phonetic distance from critical-band spectra: a first step". *Proc. IEEE Inf. Conf. Acoust. , Speech, and Signal Process. , pp. 1278-1281, 1982.*

[11] H. Wakita. "Linear prediction voice synthesizers: Line spectrum pairs (LPP) is the newest of several techniques". *Speech Technol. , pp. 17-22, Fall 1981.*

[12] J. Lalou. "The information index: an objective measure of speech transmission performance". *Ann. Telecommun. , vol. 45, pp. 47-65, 1990.*

[13] T. Saadawi, A. Ammar, and A. Elhakeem. "Fundamentals of Telecommunications Networks". *John Wiley and Sons, 1994 .*

[14] Peter Kroon. *,Bell Laboratories personal communication.*

[15] P. Lindsay, and D. A. Norman. "Human information processing: An introduction to psychology". *Academic Press, New York, second printing, 1972 .*

[16] E. Zwicker, and U. Zwicker. "Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system". *J. Audio Eng. Soc. vol. 39, pp. 115-125, March 1991.*

[17] J. Allen. "Cochlear modelling". *IEEE ASSP Mag., vol 2, pp. 3-29, January 1985.*

[18] D. Sen, D. H. Irving, W. H. Holmes. "Use of an auditory model to improve speech coders". *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing, Minnesota, April 27-30, 1993.*

[19] R. A. W. Bladon, B. Lindblom. "Modling vowel perception". *J. Acoust. Soc. Am. , vol. 69, pp. 1414-1422, May 1981.*

[20] A. Fourcin. "Speech processing by man and machine-Group report". *Recognition of Complex Acoustic Signals, T. Bullock, Ed. Life Sciences Res. Rep. 5 of the Dahlem Workshops, Berlin, Germany, 1977.*

[21] R. E. P. Dowling, and L. F. Turner. "Modeling the detectability of changes in auditory signals". *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing, Minnesota, April 27-30, 1993.*

[22] E. Zwicker, and H. Fastl. "Psychoacoustics: Facts and Models". *Springer-Verlag, Berlin, 1990.*

[23] H. Hermansky. "Perceptual linear predictive (PLP) analysis of speech". *J. Acoust. Soc. Am. , vol. 87, pp. 1738-1752, April 1990.*

[24] W. D. Robinson, and R. S. Dadson. "A redetermination of the equal-loudness relationships for pure tones". *Br. J. Appl. phys. 7, pp. 166-181, 1956.*

[25] P. Hess, and G. J. Montgomery. "Abduction: Theory and Application". *proceedings of the 4th Annual Aerospace Applications of Artificial Intelligence Conference (AAAIC), Oct. 1988.*

[26] G. J. Montgomery. "Abductive Diagnostics". *proceedings of AIAA Computers in Aerospace Conference, Monterey, CA, pp. 267-275, Oct. 3-5, 1989.*

[27] P. Hess. "Neural Network Approach to Problem Dealing with Uncertainty". *proceedings of the 3th Annual Aerospace Application of Artificial Intelligence Conference (AAAIC) pp. 89-100 , October 1987.*

[28] R. L. Barron, et al. "Adaptive Learning Networks: Development and Application in the United States of Algorithms Related to GMDH". *Self-Organizing Methods in*

Modeling; GMDH Type Algorithms, edited by S. J. Farlow, Marcel-Dekker, Inc., New York, 1984.

[29] S. Stefan, R. L. Barron, and L. O. Gilstrap. "Polynomial and Neural Networks: Analogies and Engineering Applications". *IEEE International Conference on Neural Networks, San Diego, California, June 20-24, 1987.*

[30] A. R. Barron. "Predicted Squared Error: A Criterion for Automatic Model Selection". *Self-Organizing Methods in Modeling: GMDH Type Algorithms, edited by S. J. Farlow, Marcel-Dekker, Inc. , New York, 1984.*

[31] Demonstration Package, AbTech Corporation. .

[32] G. A. Miller. "The Magic Number Seven, Pluse or Minus Two: Some Limits on our Capacity for Processing Information". *The Psychological Review, 63, pp. 81-97, March 1956.*

[33] R. Kubichek, D. Atkinson, A. Webster. "Advances in objective quality assessment". *GLOBCOM Technical Proceedings, Phoenix, AZ, Dec. 2-5, 1991.*

[34] E. Zwicker, E. Terhardt. "Analytical expressions for critical-band rate and critical bamdwidth as a function of frequency". *J. Acoust. Soc. Am. , vol. 68, no. 5, pp. 1523-1525, Nov. 1980.*

[35] S. Michaelson and M. Steedman. "Hidden Markov Models for Speech Recognition". *X. D. Huagn, Y. Ariki and M. A. Jack , pp. 41, 1990.*

[36] S. Chen, C. F. N. Cowan, and P. M. Grant. "Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks". *IEEE Transactions on Neural Networks, vol. 2, no. 2, pp. 302-309, March 1991.*

[37] J. Park and I. W. Sandberg. "Universal approximation using radial-basis function networks". *Neural Computation, vol. 3, pp. 246-257, 1991.*

[38] J. Moody and C. J. Darken. "Fast learning in networks of locally tuned processing units". *Neural Computation, vol. 1, pp. 281-294, 1989.*

[39] J. A. Leonard, M. A. Kramer, and L. H. Ungar. "Using Radial Basis Functions to Approximate a Function and Its Error Bounds". *IEEE Transactions on Neural Networks, vol. 3, no. 4, pp. 624-627, July 1992.*

[40] Matlab. Neural Network Toolbox. *ch. 6, pp. 1-12.*

[41] M. de Pryker. "Asynchronous Transfer Mode: Solution for Broadband ISDN ". *New York: Ellis Horwood, second edition, 1993.*

[42] H. Gilbert, O. A. Magd, and V. Phung. " Developing a cohesive Traffic Management Strategy for ATM Networks". .

[43] K. Kondo, and M. Ohno. " Packet Speech Transmission on ATM Networks using a variable Rate Embedded ADPCM Coding Scheme". *IEEE Trans. commun., vol. E76-B, no. 4, April, 1993.*

[44] I. Cidon, A. Khamisy, and M. Sidi. " Dispersed Messages in Discrete-Time Queues: Delay, Jitter and Threshold Crossing". *Proc. IEEE INFCOM'94, Toronto, Ontario , Canada, pp. 218-223, June 12-16, 1994.*

[45] N. Jayant. "High Quality Networking of Audio-Visual Information ". *IEEE commun. Magazine, pp. 84-95, 1993.*

[46] D. D. Clark, S. Shenker, L. Shang. " Supporting real-time applications in an integrated services packet network: Architecture and mechanism". *Proc. ACM Sigcomm'92, Baltimore, MD, pp. 14-26, Aug. 1992.*

[47] K. W. Gould, et. al. "Robust Speech Coding for the Indoor Wireless Channel ". *AT and T Technical J. , Oct-Nove. , 1993.*

[48] L. A. Aurel and P. Giovanni. " Control of resources in broadband networks with quality of service guarantees". *IEEE Commun. Mag. vol. 29, pp. 66-73, Oct. 1991.*

[49] Anagnostou, M. E, Theologou, M. E, Vlakos, K. M , Tournis, D and Protonotarios, E. N. " Quality of service requirements in ATM based B-ISDN". *Comput. Commun. , vol. 14, no. 4, pp. 197-204, May, 1991 .*

[50] W. Matragi, C. Bisdikian, and K. Sohraby. " Jitter Calculus in ATM networks: Single Node Case". *Proc. IEEE INFCOM'94, Toronto, Ont. , Canada, pp. 232-241, June 12-16, 1994.*

[51] I. Cidon, A. Khamisy, and M. Sidi. " Dispersed Messages in Discrete-time Queues: Delay, Jitter and Thredhold Crossing". *Proc. IEEE INFCOM'94, Toronto, Ont. , Canada, pp. 218-223, June 12-16, 1994.*

[52] Domenico Ferrari. " Delay jitter control scheme for packet-switching internetworks". *Computer Communications, vol. 15, no. 6, pp. 367-373, July/Augst. , 1992.*

[53] J. D. Russell. " Multimedia Networking Performance Requirements". *IBM Research Report.*

[54] M. Meky, T. N. Saadawi. " Degradation Effect of Cell Loss on Speech Quality Over ATM Network". *International IFIP-IEEE Conference on Broadband Communications, Montreal, Quebec, Canada, April 23-25, 1996.*

[55] N. Jayant. " High Quality Networking of Audio-Visual Information". *IEEE commun. Mag. , pp. 84-95, 1993.*

[56] D. D. Clark, S. Shenker, L. Shang. " Supporting real-time applications in an integrated services packet network: Architecture and mechanism". *Proc. ACM Sigcomm'92, Baltimore, MD, pp. 14-26, Aug. 1992.*

[57] M. Meky, T. N. Saadawi. " A Perceptually-Based Objective Measure for Speech Coders using Abductive Network". *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, Georgia, May 7-10, 1996.*

[58] M. Meky, T. N. Saadawi. " Prediction of Speech Quality Using Radial Basis Functions Neural Networks". *ATIRP: Advanced Telecommunications/Information Distribution Research Program, Maryland, January 21-22, 1997.*

[59] W. A. Montgomery. " Techniques for Packet Voice Synchronization". *IEEE JSAC, vol. SAC-1. no. 6, pp. 1022-1027, December 1983.*

[60] F. Guiliemin and J. W. Roberts. " Jitter and Bandwidth Enforcement". *Proceeding of GLOBECOM, Phoenix, AZ, Dec. 2-5, 1991.*

[61] M. Meky, T. N. Saadawi. "Impact of Delay Jitter on Speech Quality Over ATM Networks". *ATIRP: Advanced Telecommunications/Information Distribution Research Program, Maryland, February 5-6, 1998.*

[62] L. R. Rabiner, and R. W. Schafer. "Digital Processing of Speech Signals ". *Prentice Hall, 1978.*

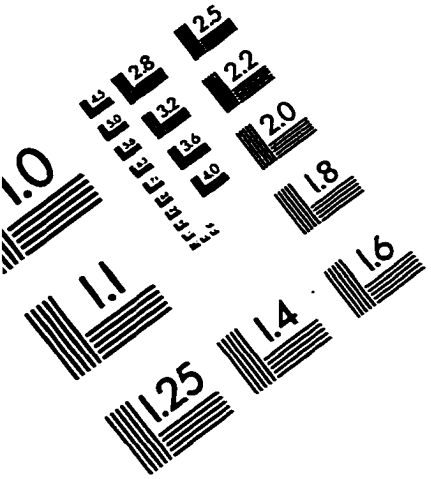
[63] R. W. Schafer, and L. R. Rabiner. " System for automatic formant analysis of voiced speech". *J. Acoustic Soc. America, vol. 47, pp. 634-648, 1970.*

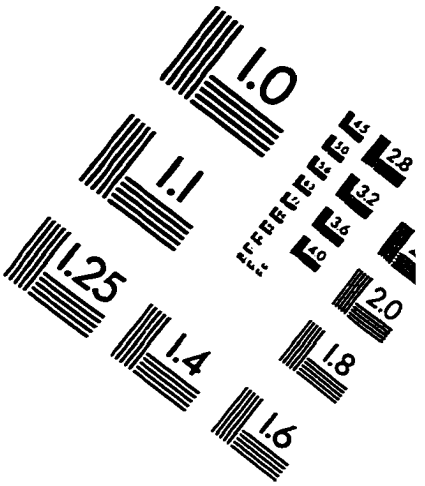
[64] B. S. Atal. " Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". *J. Acoustic Soc. America, vol. 55, pp. 1304-1312, 1974.*

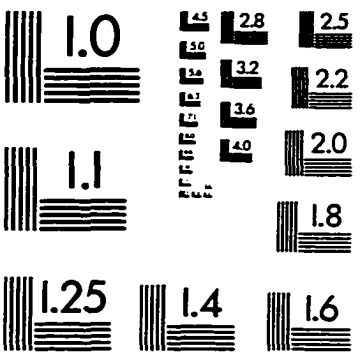
[65] A. V. Oppenheim, and R. W. Schafer. " Homomorphic analysis of speech". *IEEE Trans. Audio, Electroacoust., vol. AU-16, pp. 221-226, 1968.*

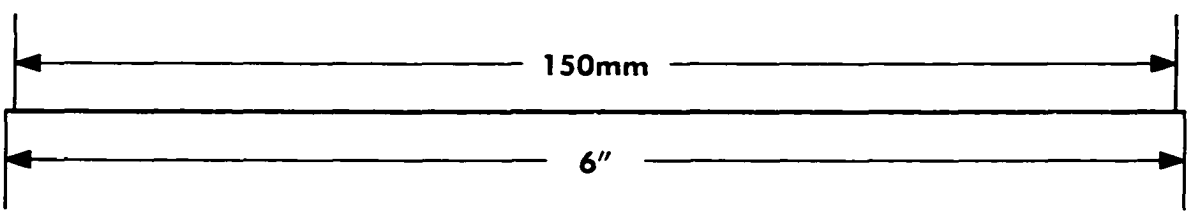
[66] L. R. Rabiner, J. G. Wilpon, and F. K. Soong. "High performance connected digit recognition using hidden Markov modles ". *Proc. ICASSP-88, New York, 1988.*

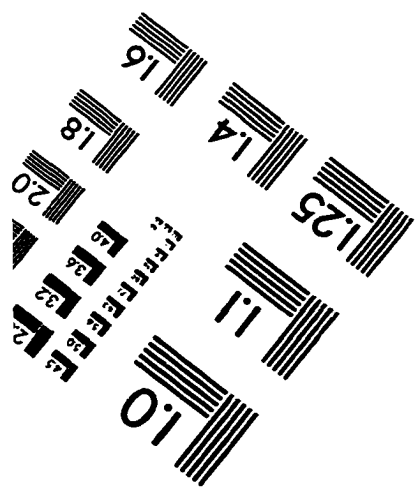
IMAGE EVALUATION TEST TARGET (QA-3)

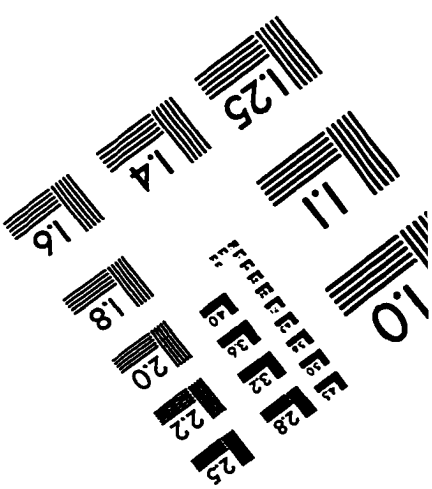












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989