Research on Objective Speech Quality Measures

by

Carol S. Chow

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of
Bachelor of Science in Electrical Engineering and Computer Science
and Master of Engineering in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 2001

Author_____
Department of Electrical Engineering and Computer Science
February 1, 2001

Certified by_
Vishu R. Viswanathan
VI-A Company Thesis Supervisor

Certified by_
Thomas F. Quatieri
M.I.T. Thesis Supervisor

Accepted by_____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Research on Objective Speech Quality Measures
by
Carol S. Chow

Submitted to the
Department of Electrical Engineering and Computer Science

February 6, 2001

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Electrical Engineering and Computer Science
and Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

This is a thesis dissertation on objective speech quality measures. Two objective measures, Enhanced Modified Bark Spectral Distortion (EMBSD) and Perceptual Evaluation of Speech Quality (PESQ) were included in this study. The scope of the study covers the evaluation of EMBSD and PESQ in predicting subjective results from Mean Opinion Score (MOS) tests; an extension of PESQ to handle wideband speech; and the performance of EMBSD and PESQ on Degradation Mean Opinion Score (DMOS) tests in noise conditions. The following results are reported: (1) EMBSD can predict the quality of various conditions for a given coder, but not across coders. (2) PESQ can predict the quality of various conditions for a given coder as well as across coders. (3) While PESQ is effective in handling time shifts that occur during silence, it does not seem as effective when such shifts occur during speech. (4) A simple extension of PESQ can evaluate wideband speech as well as it evaluates narrowband speech. (5) When clean speech is used as reference, EMBSD predicts DMOS better than when noisy speech is used as reference. (6) PESQ predicts DMOS better when using noisy speech than with using clean speech as reference.

Thesis Supervisor: Vishu R. Viswanathan
Title: TI Fellow, Speech Coding R&D Manager, DSP R&D Center, Texas Instruments

Thesis Supervisor: Thomas F. Quatieri
Title: Senior Member of the Technical Staff, M.I.T. Lincoln Laboratory

Acknowledgments

I would like to thank Texas Instruments for sponsoring me through the VI-A Program. I would also like to thank Vishu Viswanathan for his guidance, his time, and his efforts in helping me during this thesis work. Many thanks to the following: Antony Rix of British Telecom, who helped explained PESQ and participated in discussions; Professor Robert Yantorno of Temple University, who provided EMBSD and gave helpful comments; and John Tardelli of Arcon Corporation, who provided speech and MOS databases T7 and T8. I would like to thank the members of the Speech Coding Branch: Wai-Ming Lai, Anand Anandakumar, Alexis Bernard, Alan McCree, Erdal Pakstoy, and C. S. Ramlingham. I owe much gratitude to Őzge Nadia Gőzűm, my brother Allan, Barney Ramirez, and Richard Monté for being there when I needed a break. Most importantly, I would like to thank my parents who were always just a phone call away.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Speech quality assessment is an essential part of the development of speech coders. Effective speech quality measures make it possible to evaluate speech coders during development, to compare different speech coders, and to measure the quality of speech communication channels.

Since speech quality is ultimately judged by the perception of speech by human listeners, current speech quality tests are performed mainly by subjective measures that use human listeners to evaluate speech samples. However, subjective tests are time consuming, costly, and not highly consistent. These disadvantages have motivated the development of objective measures that can predict subjective scores, but without using human listeners.

This thesis is on objective speech quality measures. The scope of the thesis includes the evaluations of two objective measures, Enhanced Modified Bark Spectral Distortion (EMBSD) and Perceptual Evaluation of Speech Quality (PESQ), and attempts to use objective measures for wideband speech evaluation and for the prediction of Degradation Mean Opinion Scores. The study of PESQ was performed as part of the effort at Texas Instruments to evaluate PESQ for the ITU-T standardization process.

The major findings of this thesis are as follows:

- EMBSD accurately predicts the quality of various conditions for a given coder, but does not consistently predict the quality across coders. (Chapter 4)
- The feasibility of using forward masking in the frame distortion measure of EMBSD was investigated and found to be unsatisfactory. (Chapter 4)
- The performance of L1 averaging was compared with the performance of Peak Over Block (POB) averaging. Although L1 averaging is better under certain conditions, POB performs better overall. (Chapter 4)
- PESQ accurately predicts the quality of various conditions for a given coder as well as across coders. (Chapter 5)
- The time alignment mechanism in PESQ is effective in handling time shifts that occur during silent periods. However, it does not seem as effective when such time shifts occur during speech periods. (Chapter 5)
- A simple extension of PESQ, denoted as PESQ-WB, evaluates wideband speech as well as PESQ does on narrowband speech. (Chapter 6)
- The ability of EMBSD and PESQ in predicting DMOS in noise conditions was investigated. When clean speech is used as reference, EMBSD predicts DMOS better than when noisy speech is used as reference. On the other hand, PESQ predicts DMOS better when using noisy speech than with using clean speech as reference.

The organization of the thesis is as follows:

- Chapter 2 provides background information and describes several subjective and objectives measures.

- Chapter 3 describes the use of two metrics for evaluating objective measures: correlation coefficient between objective and subjective scores and root mean-squared error in the prediction of subjective scores.
- Chapter 4 treats the evaluation of EMBSD and investigation of it's the performance when forward masking and L1 averaging techniques are used.
- Chapter 5 focuses on the performance of PESQ, including the effectiveness of its time alignment mechanism.
- Chapter 6 introduces PESQ-WB and discusses its performance.
- Chapter 7 examines the feasibility of EMBSD and PESQ in evaluating DMOS.
- Chapter 8 presents conclusions and recommendations for further research.

Chapter 2

BACKGROUND

Since speech quality is ultimately judged by the human ear, subjective measures provide the most direct form of evaluation. Even though subjective measures evaluate speech quality in a direct manner, they have disadvantages. Subjective measures require special testing environments, human listeners, money, and time. Test scores are also inherently subjective and difficult to reproduce. These disadvantages have motivated the development of measures to predict subjective scores objectively, without human listeners. This section provides background information on subjective and objective measures that have been developed.

2.1 Subjective Speech Quality Measures

Subjective measures are based on ratings given by human listeners on speech samples. The ratings use a specified score table, and then are statistically analyzed to form an overall quality score. Discussed below are the three subjective measures that are included in the ITU-T Recommendation P.800: the Absolute Category Rating (ACR), the Degradation Category Rating (DCR), and the Comparison Category Rating (CCR) [1].

2.1.1 Absolute Category Rating

The Absolute Category Rating (ACR) produces the widely used mean opinion score (MOS). Test participants give MOS ratings by listening only to the speech under

test,without a reference. The five-point MOS rating scale is shown in Table 1.

MOS Rating	Speech Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Unsatisfactory

Table 1 MOS Rating Scale

The ACR provides a flexible scoring system because listeners are able to make their own judgment on speech quality. However, this flexibility can result in varying quality scales due to different individual preferences.

2.1.2 Degradation Category Rating

The Degradation Category Rating (DCR) measure is a comparison test that produces Degradation MOS (DMOS). In DCR, listeners compare the distorted speech with the reference speech. The reference is always played first and the listener is aware of this. The listeners use the impairment grading scale shown in Table 2 to evaluate the difference between distorted and reference signals.

DMOS Rating	Level of Distortion
5	Imperceptible
4	Just Perceptible, but not annoying
3	Perceptible and slightly annoying
2	Annoying, but not objectionable
1	Very Annoying and objectionable

Table 2 DMOS Rating Scale

DCR is often used to judge speech quality in background noise conditions such as car, street, and interference talker noise. The amount of noise and the type of noise will affect the perceived degradation level.

The format of the DMOS measure is similar to the structure of most objective measures. Therefore, some believe objective measures are better suited to predicting DMOS than to predicting MOS. More discussion is included in Chapter 7.

2.1.3 Comparison Category Rating

The Comparison Category Rating (CCR) method is another comparison test that produces Comparison MOS (CMOS). The CCR method is similar to DCR except that the distorted and reference signals are played in a random order and listener is not told which signal is the reference. The listener ranks the second signal against the first on a scale shown in Table 3.

CMOS Rating	Comparison Level
3	Much Better
2	Better
1	Slighly Better
0	About the Same
-1	Slightly Worse
-2	Worse
-3	Much Worse

Table 3 CMOS Rating Scale

If the order of the signals played is 1. Distorted 2. Reference, the raw score is reversed (i.e. –1 → 1, -2 →2, …, 2 →-2, 3→ -3) [1].

The CCR method allows the processed signal to be ranked better than the reference. Consequently, coders with characteristics such as noise suppression and signal enhancement can be rated higher than the reference. CMOS is also suitable for comparing signals coded by different coders where the better coder is not known in advance.

2.2 Objective Speech Quality Measures

There are many advantages to objective measures. Since the measures are computer based, they provide automated and consistent results. Objective measures can speed up speech coder development by automating design parameter optimization. Objective measures also do not have the disadvantages of subjective testing caused by listener fatigue and lack of concentration.

Objective measures are also useful in applications where subjective tests are ineffective. For example, Voice over Internet Protocol (VoIP) network monitoring systems can use objective measures to provide real-time feedback. A speech signal can be passed through the network and returned to the same location such that both the original and processed signals can be input into an objective measurement device. The score produced by the objective measure can report the speech quality provided by the system and immediate modifications can be made as necessary. Using subjective tests does not make sense in such real-time applications.

Most objective speech quality measures compare the distorted signal to a reference. Objective measures lack an internal model of quality and therefore, use the original, undistorted signal as the reference. There are objective measures that do not

utilize a reference signal such as the Output-Based Speech Quality, but they are not included in this research [2]. This section outlines how most objective measures compare the distorted signal to the reference, and describes the different types of objective measures.

2.2.1 Framework of Objective Measures

There is an agreement on a basic structure to design objective measures [3][4]. Figure 1 shows the structure consisting of three stages: alignment, frame distortion measure, and time averaging. Unless said otherwise (see Chapter 7), original, clean speech is used as the reference signal.

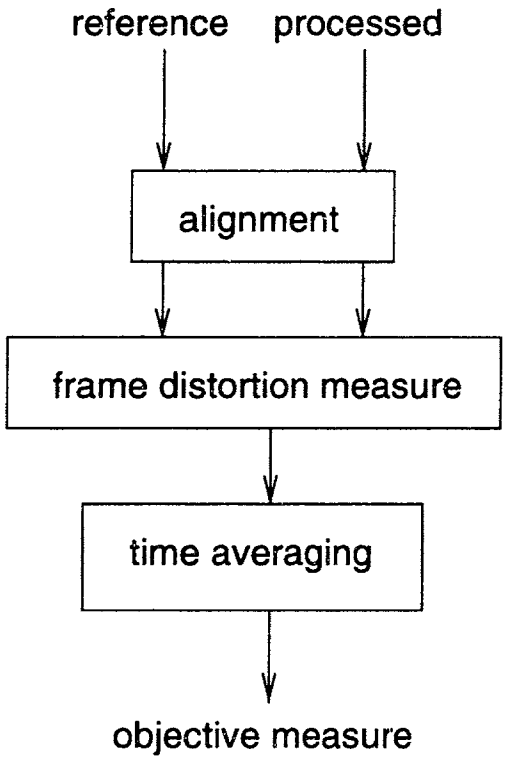

Figure 1 Objective Measure Framework

Alignment

In the alignment stage, the reference and distorted signals are compared, and time synchronization of the two signals is performed. If the signals are not time-synchronized, a large error may be erroneously calculated. Level normalization and equalization are performed, also as part of this stage. Equalization adjusts for linear filtering effects on the distorted signal relative to the reference.

Frame Distortion Measure

In the frame distortion measure stage, the speech signals are broken into short segments, or frames, with a typical duration of 10 to 30 ms. For each frame, a distortion value is calculated by comparing the distorted speech signal with the reference. The comparison may be done in time domain, frequency or spectral domain, or perceived loudness domain. Loudness domain approaches have achieved the greatest success.

Time Averaging

In the time averaging stage, frame distortions are averaged over the duration of speech under test, to produce a single overall distortion measure. An example averaging method is the Lp norm. Various weighting methods are usually incorporated to handle different types of distortions. The overall distortion measure may be mapped to produce a subjective score prediction.

2.2.2 Types of Objective Measures

Objective measures can be divided into three types: time domain, frequency domain, and perceptual measures [3]. Examples of each type are described in this section. Most of the measures encompass only the frame distortion and time averaging stages of the objective measure framework. Of the measures presented here, only Perceptual Analysis Measurement System (PAMS) and Perceptual Evaluation of Speech Quality (PESQ) include all three stages.

Time Domain Measures

Signal to Noise Ratio (SNR): Signal to Noise Ratio measures are suited for measuring analog and waveform coding systems. An SNR measure is easy to implement; however, it is very sensitive to the time alignment of the original and distorted speech signals. SNR measures compare the distorted and reference signals on a sample-by-sample basis and hence are appropriate in general for high bit-rate coders. In particular, they are not able to estimate accurately the perceived quality of low rate coders.

Segmental SNR (SNRseg): The segmental SNR measure is an improvement on the SNR. The SNRseg is an average of the SNR over smaller segments of the speech signal. The overall speech signal is broken down into smaller segments, allowing the SNRseg to achieve a greater level of granularity. Like the SNR, the usefulness of SNRseg is limited to waveform coders [3].

Frequency Domain Measures

There are a number of ways to calculate frequency domain measures. Three such measures are discussed here.

Log Likelihood Ratio (LLR): The LLR is also known as the Itakura distance measure [5]. The LLR is the distance between the all-pole model representations of the reference and distorted speech signals. The measure is based on the assumption that a p^{th} order all-pole model can represent a frame of speech. Therefore, the LLR is limited to speech signals that are well represented by an all-pole model.

Linear Prediction Coefficients (LPC): The LPC measure is based on the parameterizations of the linear prediction vocal tract models. The parameters can be prediction coefficients or transformations of the prediction coefficients. Each type of parameters quantifies the distance between the reference and distorted signal differently. Of all parameters, the log area ratios had been recorded as the best [3][6].

Cepstral Distance Measure: The cepstral distance measure is based on cepstral coefficients calculated from linear prediction coefficients. The resulting cepstrum is an estimate of the smoothed speech spectrum.

Perceptual Domain Measures

Recent objective measures have shown large improvements over time and frequency domain measures by incorporating psychoacoustic principles. These principles include critical band frequency analysis, absolute hearing thresholds, and masking.

Critical band frequency analysis helps explain how the ear processes signals. A frequency-to-place transformation takes place in the inner ear. Distinct regions in the inner ear are sensitive to different frequency bands, or critical bands. By separating signals into critical bands, objective measures can capture the particular sensitivities that the ear has to different frequencies. The absolute hearing threshold is a level that determines the amount of energy needed in a pure tone that can be detected by a listener in a noiseless environment [7]. This is used as the minimum audible threshold at which distortions must exceed in order to be considered. Masking refers to the process where one sound is made inaudible because of the presence of other sounds. Simultaneous masking refers to a frequency domain masking that is observed with critical bands. The presence of a strong masker creates an excitation in the inner ear to block the detection of a weaker signal. Nonsimultaneous masking is the extension of simultaneous masking in time. Effectively, a masker of finite duration masks signals prior to the onset of the masker (backward masking) and immediately following the masker (forward masking). Objective measures can use masking to increase the audible threshold.

Bark Spectral Distortion (BSD): The Bark Spectral Distortion (BSD) measure was developed at the University of California at Santa Barbara [8]. BSD was one of the first measures to incorporate psychoacoustic responses into an objective measure. BSD

transforms the reference and distorted signals to Bark spectral representation. The objective score is then the distance measure between the two spectra. The objective scores correlated so well with subjective scores that BSD became the basis for many new objective measures.

The BSD requires that the reference and distorted signals must be time aligned first. Once the speech signals are broken into frames, both the reference and distorted signals are transformed using psychoacoustic principles: critical band filtering, perceptual weighting of spectral energy, and subjective loudness. The method is described below.

Critical band filtering is based on the observation that the human auditory system has poorer discrimination at high frequencies than at low frequencies. The frequency axis is scaled from Hertz, f, to Bark, b using Equation 1.

$$Y(b) = f = 600\sinh(b/6) \tag{1}$$

which has been called the critical band density. A prototype critical-band filter smears the Y(b) function to create the excitation pattern, D(b). The critical-band filters are represented by F (b) in Equation 2.

$$10\log_{10} F(b) = 7 - 7.5*(b-0.215) - 17.5[0.196 + (b-0.215)^2]^{\frac{1}{2}} \tag{2}$$

The smearing operation is a straightforward convolution since all critical-band filters are shaped identically. The resulting operation is a convolution as shown in Equation 3.

$$D(b) = F(b) * Y(b) \tag{3}$$

Perceptual weighting adjusts for the fact that the ear is not equally sensitive to stimulations at different frequencies. In order to transform intensity levels at different frequencies to equal perceptual loudness levels, intensity levels are mapped against the standardized reference level set at the threshold at 1 kHz. The scale is the sound pressure level (SPL) and is measured in phons. Using equal loudness functions at 1 kHz, equation 3 converts dB intensity levels to loudness levels in phons. D(b) is the loudness intensity function in phons.

Subjective loudness deals with the perceptual nonlinearity. The increase in phons needs an adjustment in the subjective loudness. The adjustment varies with the loudness level. For example, while an increase of 10 phons is required to double the subjective loudness at 40 phons, an increase of 10 phons near threshold level increases the subjective loudness by ten times. The following equation is used to convert each phon P in D(b) to a sone subjective loudness level L.

$$L = 2^{(P-40)/10} \quad if \quad P \geq 40 \tag{4}$$

$$L = (P/40)^{2.642} \quad if \quad P < 40 \tag{5}$$

P is the phon loudness level.

The BSD score is an average across all BSD^k, where k represents the speech frame. For each segment, BSD^k is calculated with Equation 6.

$$BSD^k = \sum_{i=1}^{N} [L_x^{(k)}(i) - L_y^{(k)}(i)]^2 \quad (6)$$

x=reference, y=distorted signal
N = number of critical bands

The BSD^k are then time averaged with an Lp norm.

Enhanced Modified BSD (EMBSD): The Enhanced Modified BSD (EMBSD), developed at Temple University, is an improvement on the BSD measure [9]. A noise-masking threshold (NMT) and a peak-over-block (POB) averaging model were the improvements in EMBSD. The NMT sets a minimum intensity level. Distortions must be above this level order to be included in the distortion measure. The NMT is determined by the critical band spectrum of the reference signal, the spectral flatness measure (SFM), and the absolute hearing threshold.

The critical band spectrum produces tone-masking noises and noise-masking tones. Tone-masking noises are estimated as (14.5 + b) dB below the critical spectrum in dB, where b is the bark frequency. Noise-masking tone is estimated as 5.5 dB below the critical spectrum [10]. The SFM is used to determine if the critical band spectrum is a noise or tone.

The POB method groups consecutive frames together in sets of 10 to form a 'cognizable segment.' The maximum frame distortion value over the cognizable segment is chosen as the perceptual distortion value, P(j). A residual distortion value Q(j) is the distortion value of the previous cognizable segment scaled down by 0.8. The distortion value of the current cognizable segment is defined as the larger value, P(j) or Q(j).

Therefore, larger errors are emphasized and are allowed to mask smaller errors. The following equations summarize the process:

$$P(j) = \max(frdist(i), frdist(i-1),..., frdist(i-9)) \quad (7)$$

$$Q(j) = 0.8 * C(j-1) \quad (8)$$

$$C(j) = \max(P(j), Q(j)) \quad (9)$$

j refers to a cognizable segment.
frdist(i) is the frame distortion frame i.
i denotes the last frame in the cognizable segment.

The final EMBSD score is the average of C(j) over all j. Professor Robert Yantorno from Temple University provided the source code of EMBSD for use in this research.

Perceptual Analysis Measurement System (PAMS): The Perceptual Analysis Measurement System (PAMS) was developed at British Telecom in 1998 [11]. PAMS utilizes the psychoacoustic principles used in BSD. To improve the time-frequency transformation used in BSD, PAMS uses a bank of linear filters. PAMS also adds an alignment stage including time and level alignments and equalization functions. These improvements lead to a better evaluation on end-to-end applications than BSD, such as telephony and network communications [11].

ITU Standards: The Perceptual Speech Quality Measure (PSQM) was developed by Beerends and Stemerdink [12]. PSQM performs similar transformations as BSD and incorporates two significant changes: characterizing asymmetry in distortions and weighting distortions differently in silence and during speech. PSQM seeks to capture

the asymmetry in distortions by treating additive and subtractive distortions differently. Because additive distortions are more audible, they are weighted more heavily. Distortions that occur during speech are also more disturbing than those in silent periods. PSQM uses a weighting function to treat the two distortion types differently [12].

The I.T.U. Technology Standardization Sector performed a study during 1993-1996 on five different objective measures, one of which was the PSQM. PSQM was determined to be the best and was accepted in 1996 as the ITU-T Recommendation P. 861 for the objective measurement of narrowband speech codecs [14]. The ITU is currently in the process of replacing the P.861 with a new recommendation in 2001. PSQM had limitations; it could not reliably evaluate channel error conditions. The draft of the ITU-T P.862 recommendation introduces a new objective measure, the Perceptual Evaluation of Speech Quality (PESQ) [13]. PESQ is a combination of both the PSQM and the PAMS.

Perceptual Evaluation of Speech Quality (PESQ): PESQ overcomes many of the limitations faced by previous measures, such as linear filtering and delay variations. Linear filters may not have much effect on subjective quality, but it can cause the distorted signal to be very different from the reference. PESQ applies filters to equalize the distorted signal to the reference in order to avoid evaluating inaudible differences as errors. PESQ also improves upon the time alignment capability of PAMS. The time alignment component in PESQ tries to resolve time misalignments in silent periods as well as speech periods. The PESQ measure was provided to Texas Instruments (TI) for the purpose of evaluation as part of the ITU-T recommendation process.

Chapter 3

EVALUATING OBJECTIVE MEAURES

The performance of an objective measure is assessed by comparing its scores to the subjective measure it tries to predict. In PESQ and EMBSD measures, objective measure scores are compared to MOS. The process of evaluating objective measures begins with obtaining reference databases. The database contains sentence pairs that are phonetically balanced and spoken by males and females. As shown in Figure 2, the reference database is then processed to form the distorted database. Distortion can be coding distortions, channel errors, background noise, and time delays. A test condition may involve one or more of these distortions.

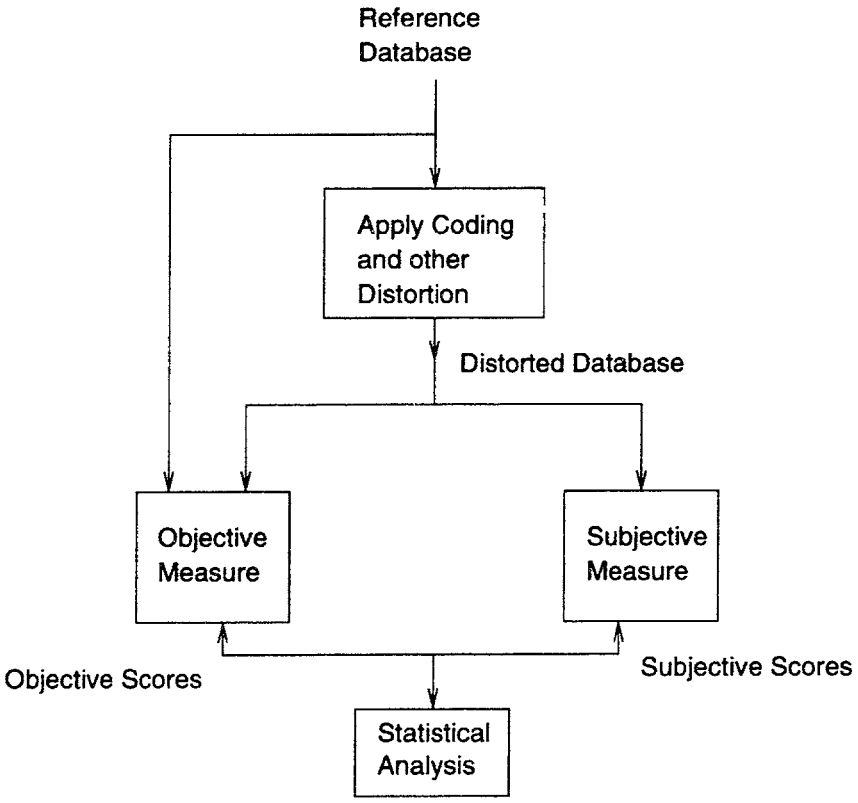

Figure 2 Process for Evaluating Objective Measures

Objective and subjective scores are collected for the entire database and the scores are averaged over all sentence pairs, for each condition. Comparisons between the objective and subjective measures will be made using the averaged condition scores only. Future references to scores will refer to the averaged condition scores.

Additional processing to "linearize" objective scores is required before they can be compared with subjective scores. Because of the nature of subjective measures, subjective ratings are affected by factors such as listener preferences and the context of a test. For example, the relative quality of the coders included in the test affects overall scores. If a mediocre quality coder A is tested with high quality coders, Coder A will score lower than if it was tested with low quality coders. For these reasons, it is difficult to directly compare two subjective tests. Some form of mapping may be necessary to compensate for these differences. The same argument applies to comparing objective scores with to subjective scores.

It is reasonable to expect the order of the conditions should be preserved, so that difference between two sets of scores should be a smooth, monotonically increasing mapping [13]. The ITU-T recommends a monotonic 3^{rd}-order polynomial function [13]. For each subjective test a separate mapping is performed on the objective scores; the mapped objective scores are then compared with the subjective scores for the test under consideration. Scores that undergo this mapping process will be referred to as polynomial-mapped scores.

The correlation coefficient between objective scores X(i) and subjective measures Y(i) is shown in Equation 10.

$$\rho = \frac{\sum_{i=1}^{N} (x(i) - \bar{x})(y(i) - \bar{y})}{\sqrt{\left(\sum_{i=1}^{N} (x(i) - \bar{x})^2 (y(i) - \bar{y})^2 \right)}} \tag{10}$$

x(i) is the i^{th} objective score.
y(i) is the i^{th} subjective score.
N is the total number of scores.

Correlation coefficients range from -1 to +1. As the value approaches +1, the two sets of data are more alike.

The correlation coefficient gives a reasonable estimate of overall similarity found in the two sets of scores. However, the metric is particularly sensitive to outliers, which can greatly improve or degrade a correlation coefficient. Also, the correlation coefficient does not take into the account the significance of differences in the two sets of scores.

The Root Mean-Squared Error metric, used along with the correlation coefficient, provides a better evaluation of the objective measure. The RMSE gives an average distance measure or error between the objective and the subjective scores, as shown in equation 11.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (X(i) - Y(i))^2}{N}} \tag{11}$$

X(i) is the i^{th} objective score.
Y is the i^{th} subjective score.
N is the total number of scores.

Before computing the RMSE measure, it makes sense to map the objective score to provide a prediction of the subjective score, as we are trying to compute the RMS prediction error.

The RMSE can also be obtained from the standard deviation of the subjective measure and the correlation coefficient as shown in Equation 12.

$$RMSE = \sigma\sqrt{(1-\rho^2)} \tag{12}$$

σ is the standard deviation of the subjective scores.
ρ is the correlation coefficient.

Given the same correlation coefficient, the RMSE will decrease as the variation in the subjective scores gets smaller.

A smaller RMSE shows that the two sets of scores are more closely related in terms of numerical value. The RMSE characterizes the prediction capability of an objective measure and should be used to evaluate this capability. As noted above, objective measures that produce distortion scores need to be mapped to provide prediction values before they can be evaluated by the RMSE metric. In assessing the RMSE, it is worth noting that MOS score differences are usually statistically significant if the differences are at least 0.15.

Chapter 4

INVESTIGATION OF EMBSD

This chapter discusses the evaluation of EMBSD and investigation of its the performance when forward masking and L1 averaging techniques are used.

4.1 Forward Masking

EMBSD uses the Peak Over Block (POB) method to implement forward masking. POB generalizes the masking across critical bands and time, and groups all masking signals in time frames together across critical bands. According to research in psychoacoustics, forward masking is frequency sensitive. A different forward masking technique is investigated to evaluate the sensitivity of masking within a critical band. Two changes are made to the EMBSD. First, the masking threshold based on simultaneous masking is replaced with a comprehensive threshold that is based on simultaneous and forward masking. Second, the POB method is replaced with a L1 norm average.

In order to capture the effect of maskers in previous frames and to allow the larger maskers to increase the audible threshold in future frames, a comprehensive threshold is used. The comprehensive threshold extends the simultaneous masking thresholds up to 200 ms, the equivalent of 10 frames. Consequently, it increases the threshold at a given particular critical band by the maximum threshold value over the set of previous 9 and current frames. A scale factor is used to reduce the threshold with each additional frame, and consequently, to decrease the effect of a masker over time.

The process is summarized in Equation 13.

$$FNMT_i[j] = \max(d^q * nmt_{(i-q)}[j],\ q = 0,1,2,...9) \qquad j = 1,2,...,B \tag{13}$$

$FNMT_i$ is the new comprehensive noise-masking threshold in sones for frame i.
nmt_i is the noise masking threshold for the frame i.
j is the critical band number
B is the total number of critical bands.
d is the scale factor.

The original EMBSD and EMBSD with the new comprehensive threshold are tested on four databases. Data are summarized in Table 4. The results show that the new comprehensive threshold did not improve EMBSD over the original masking threshold.

Database	Correlation Coefficient	
	EMBSD w/ forward masking	EMBSD
A	0.91	0.92
B	0.75	0.83
C	0.84	0.89
D	0.80	0.85

Table 4 Correlation Coefficient Data for EMBSD and EMBSD w/ forward masking

[d =0.75, d=0...0.90 were tested and all values performed worse than EMBSD]

Beerends also performed forward masking experiments. His efforts to apply forward masking to the PSQM were also not beneficial. According to Beerends, masking effects may not be applicable to telephone-band speech because of the limited bandwidth and the large distortion [16].

4.2 POB vs. L1 norm

The L1 norm method was used in an earlier version of EMBSD. It was subsequently replaced by the Peak Over Block (POB) method, to better evaluate background noise and bursty error distortions [9]. Even though POB has shown improved correlation results, it

may not be the best method in all situations. For example, the L1 norm may be more effective in 'no error' conditions or when errors are not bursty. To explore this possibility, the comparison between POB and L1 norm was studied. Let EMBSD-L1 refer to the implementation of EMBSD with L1 norm.

The average frame distortion used in EMBSD-L1 was shown in Equation 14:

$$EMBSD - Lp \quad score = \left\{ \frac{1}{N} \sum_{l=1}^{N} (frame distortion[l])^{p} \right\}^{\frac{1}{p}} \tag{14}$$

N is the total number of frames
p is the order of the Lp norm

Quackenbush, et.al. evaluated the effect of Lp averaging in various spectral distance measures [3]. They reported that variations in p had a moderate effect on the performance of objective measures. In certain cases, higher correlations were obtained for lower values of p. Lp norms other than L1 were tested for a few databases; however, correlation scores did not show consistent improvement over L1 and POB.

4.3 Performance of EMBSD

Seven test databases are used to evaluate the performance of EMBSD and EMBSD-L1. The descriptions of these databases are listed in Table 5.

Database	# of Conditions Evaluated	Coders tested	Conditions
T1	32	GSM Full Rate, GSM Enhanced Full Rate, G.728, CELP coders at 7.45 kb/s and 11.85 kb/s source coding bit rates	channel signal to interference ratios (C/I) of 19 dB to 1dB, level variations, no error, MNRU
T2	22	G.726, G. 729, 4kb/s coders	Bit error, frame erasure, no error, tandem, level variations, MNRU
T3	32	G.726, G.729, G.729D, G.723.1, 4 kb/s coders	level variations, tandem conditions, no error, MNRU
T4	32	G.726, G.729, G.729D, G.723.1, 4 kb/s coders	bit error, frame erasure, MNRU
T5	32	G.726, G.729, G.729D, G.723.1, 4 kb/s coders	level variations, tandem, no error, MNRU
T6	20	CELP coders at 11.9 kb/s & 9.5 kb/s, PCS1900 at 13 kb/s, Variable Rate CELP at 9.6 kb/s &5.8 kb/s, G.728, GSM Full Rate, GSM Half Rate	C/I 4,7, and 10 dB, no error, tandem, MNRU
T7/T8	39	CVSD at 16 kb/s, CVSD at 8 kb/s, G.726, VSELP, LPC, two STC at 2.4 kb/s, STC at 4.8 kb/s, MBE at 2.4 and 4.8 kb/s, CELP at 4.8 kb/s	bit error, no error, and jeep noise distortions, MNRU

Table 5 Database Descriptions

T7 and T8 were provided by Arcon Corporation for evaluating objective measures. Both use the same speech data and test plan and were evaluated by two different listening groups.

4.3.1 Distortion Mapping for MOS Prediction

The goal here is to map or transform distortion values to provide a prediction of the MOS score. One way to map distortion values is to use the regression line between MOS and EMBSD scores as the mapping function, as given in Equation 15.

$$\hat{y} = ax + b \qquad s.t. \quad \min_{a,b} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{15}$$

y is the MOS value,
x is the distortion value
$\hat{y}$ is the corresponding mapped value

Different databases can be used to calculate the regression used to determine the values of a and b. Two possible methods are presented. Method 1 developed at Bell Laboratories is based on the assumption that the MOS of MNRU conditions are consistent across different databases[17]. A regression analysis is performed on MNRU conditions collected from many databases. The resulting parameters are used to create the distortion-to-MOS mapping function. Method 2 seeks to create a mapping function that is computed or trained over many databases, involving coded speech and MNRU conditions. The training database is obtained by selecting roughly one-half of the conditions from the available databases.

Both methods are applied to the eight databases and the results are presented in Table 6.

Data Set	EMBSD		Polynomial Mapped EMBSD	
	Method 1	Method 2	Method 1	Method 2
T1	0.51	0.49	0.24	0.24
T2	0.69	0.67	0.21	0.21
T3	0.45	0.45	0.34	0.34
T4	0.37	0.33	0.25	0.25
T5	0.55	0.49	0.32	0.32
T6	0.63	0.74	0.52	0.52
T7	0.69	0.93	0.38	0.38
T8	0.78	0.98	0.42	0.42

Table 6 RMSE values for Methods 1 and 2.

Although EMBSD shows different results for the two methods, polynomial mapped EMBSD results are identical. Since Method 2 is better for more databases under the raw EMBSD scores, it is used for the remainder of the KMSE calculations that are presented in this dissertation.

4.3.2 EMBSD Results

EMBSD scores were computed and mapped for all eight databases. The correlation coefficients and RMSE values are listed in Table 7 and EMBSD versus MOS scatter plots are shown in Figures 3(a)-(p). In each row, the left plot shows the raw EMBSD scores; the smooth curve shown is the 3rd-order polynomial mapping function. The right plot shows the polynomial–based EMBSD.

Database	EMBSD		Polynomial Mapped EMBSD	
	Correlation	RMSE	Correlation	RMSE
T1	0.91	0.42	0.96	0.24
T2	0.92	0.67	0.95	0.21
T3	0.83	0.45	0.85	0.34
T4	0.89	0.33	0.92	0.25
T5	0.85	0.49	0.86	0.32
T6	0.70	0.74	0.72	0.52
T7	0.70	0.93	0.88	0.38
T8	0.71	0.99	0.86	0.42

Table 7 Correlation and RMSE data for EMBSD

Correlations above 0.9 for T1, T2, and T4 show that EMBSD is good at predicting speech coded by high quality coders that included many error distortions, such as bit error, frame erasure, and channel error conditions. The measure was not as effective in databases (T6-T8) where lower quality coders and no error conditions were predominant.

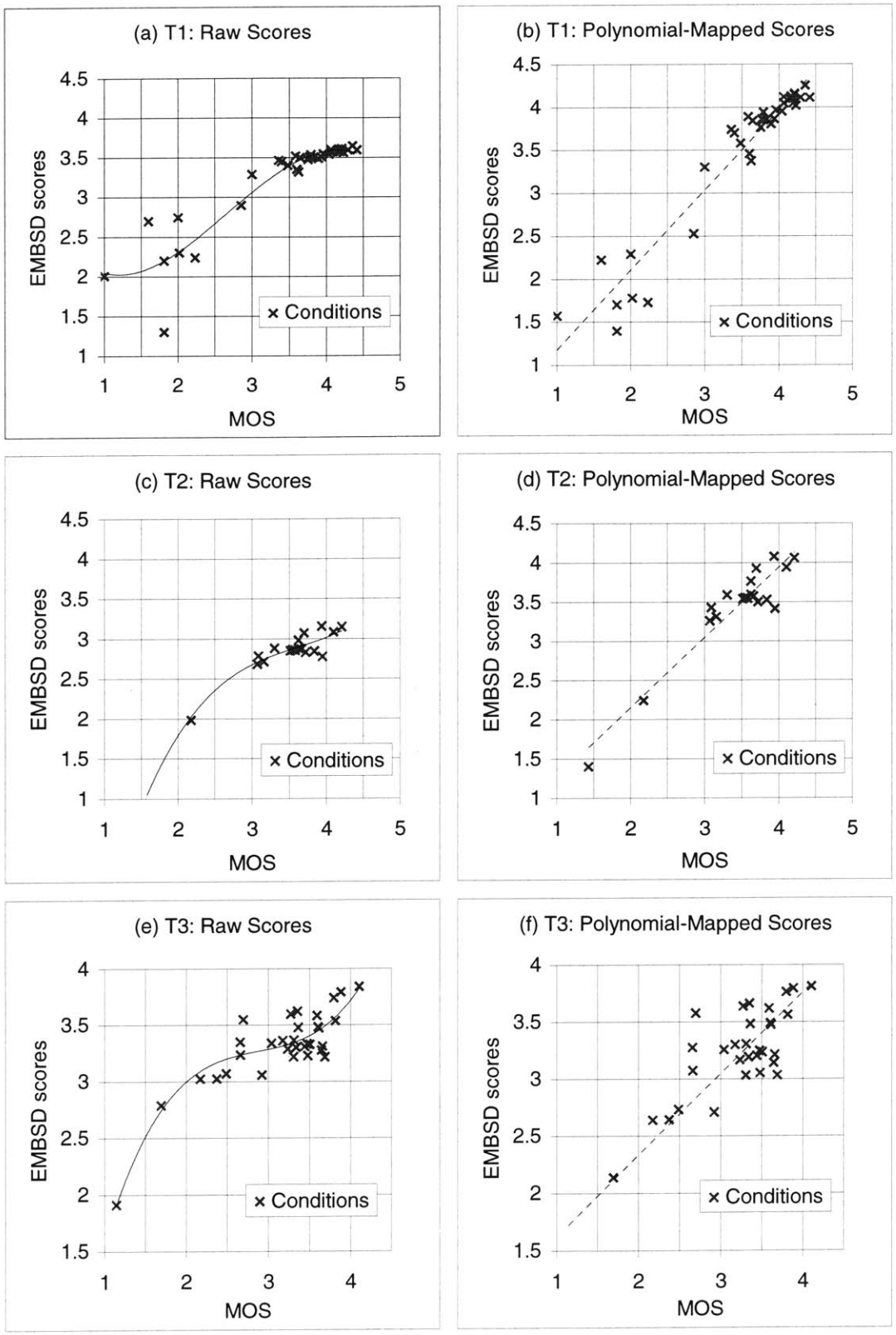

Figure 3 (a)-(f) Plots of EMBSD vs. MOS before and after polynomial-mapping

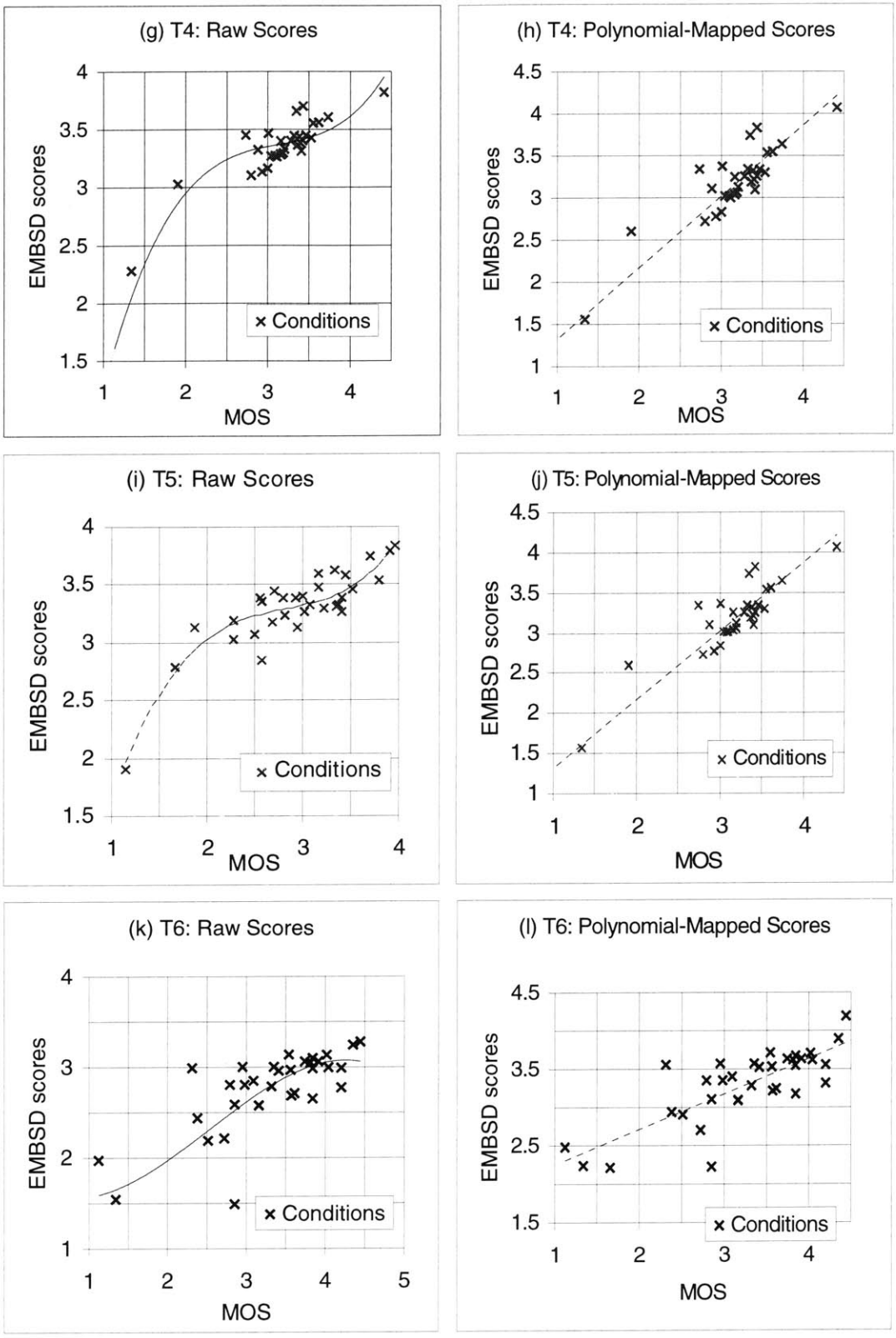

Figure 3 (g)-(l) Plots of EMBSD vs. MOS before and after polynomial-mapping

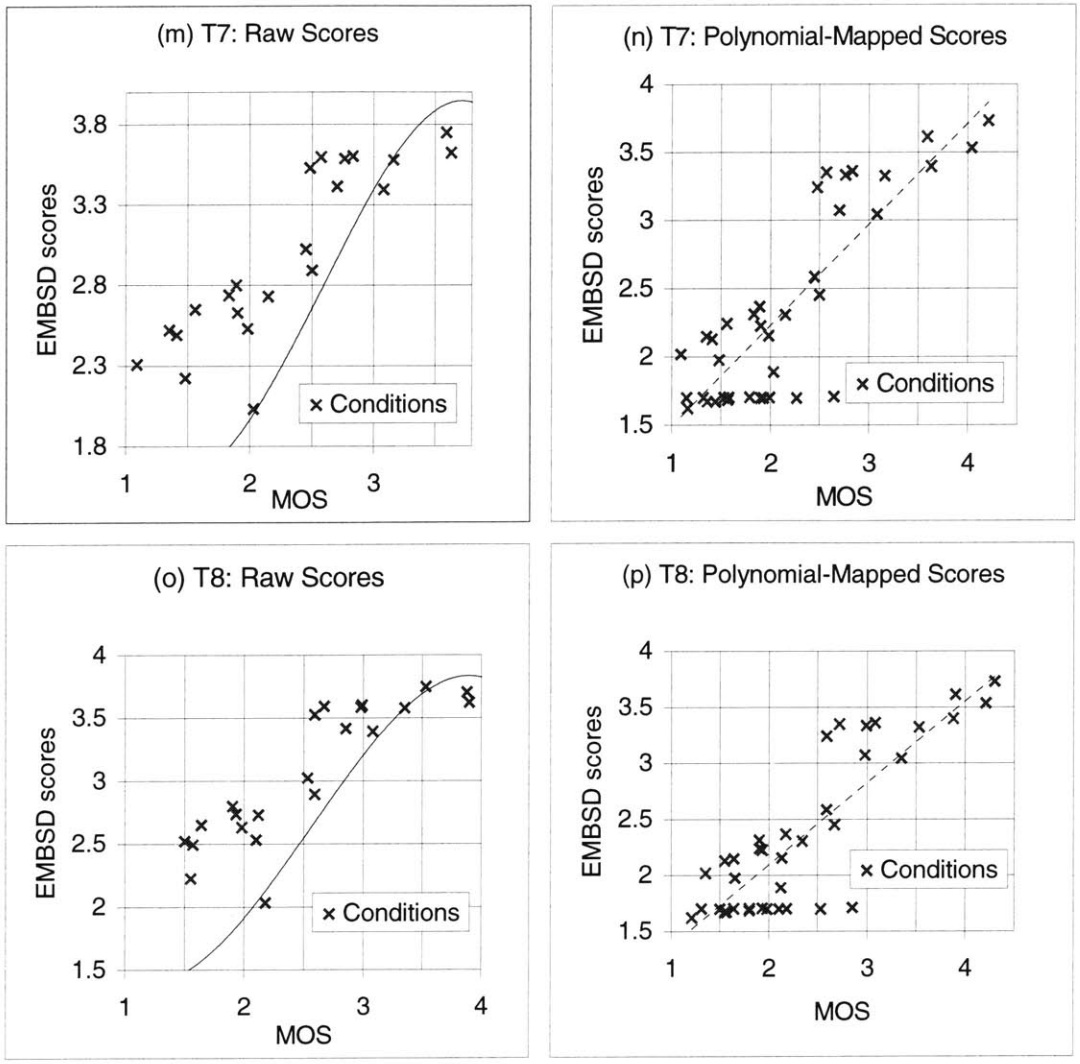

Figure 3 (m)-(p) Plots of EMBSD vs. MOS before and after polynomial-mapping

4.3.3 EMBSD-L1 Results

EMBSD-L1 scores were calculated for the eight databases. Results are shown in Table 8.

	EMBSD-L1		Polynomial mapped EMBSD-L1	
Data Set	**Correlation**	**RMSE**	**Correlation**	**RMSE**
T1	0.84	0.54	0.94	0.30
T2	0.91	0.71	0.92	0.23
T3	0.80	0.42	0.82	0.37
T4	0.85	0.66	0.96	0.17
T5	0.82	0.46	0.88	0.30
T7	0.62	1.47	0.78	0.50
T8	0.64	1.50	0.79	0.45

Table 8 Correlation and RMSE data for EMBSD-L1 before and after polynomial-mapping

Comparisons between the correlation coefficients of EMBSD in Table 7 and EMBSD-L1 in Table 8 show that EMBSD is better than EMBSD-L1 in five of the seven data sets. Though EMBSD-L1 produces improvement in T4 and T5, the improvement is small.

Next, performance was investigated over individual condition-groups including no error, error, MNRU, high-rate coders, mid-rate coders, low-rate coders, noise, and tandem. Since each condition-group has a different range of subjective ratings and a different numbers of conditions, it may be inappropriate to directly compare correlation coefficients among condition-groups [18]. Therefore, only RMSE results are used for this analysis.

Table 9 displays the comparison of RMSE for EMBSD and EMBSD-L1. The comparison is shown for no error, error, MNRU, high-rate, mid-rate, low-rate, noise, and tandem condition-groups. The error condition-group includes bit error, frame erasure, and different C/I levels. The high-rate condition-group includes G.726, CVSD, G.726, G.729, and GSM-enhanced full rate. The mid-rate condition-group includes coders such as VSELP, G.723.1, GSM full rate, and coders around 5 kb/s. The low-rate condition-group includes coders like LPC, MBE, STC, and low-rate CELP.

Category	Objective Measure	RMSE						
		T1	T2	T3	T4	T5	T7	T8
No error	EMBSD	0.79	0.25	0.18	0.28	0.34	0.34	0.30
	EMBSD-L1	0.76	0.33	0.18	0.27	0.30	0.35	0.41
Error	EMBSD	0.26	0.18	0.32	0.30	X	0.48	0.34
	EMBSD-L1	0.31	0.19	0.36	0.25	X	0.72	0.68
MNRU	EMBSD	0.30	0.14	0.34	0.55	0.23	0.42	0.59
	EMBSD-L1	0.41	0.12	0.19	0.31	0.12	0.59	0.56
High-rate	EMBSD	0.22	0.26	0.49	X	0.30	0.49	0.44
	EMBSD-L1	0.21	0.34	0.43	X	0.35	0.44	0.47
Mid-rate	EMBSD	0.34	0.23	0.31	0.15	0.33	0.61	0.79
	EMBSD-L1	0.40	0.30	0.40	0.13	0.28	0.47	0.57
Noise	EMBSD	X	X	X	X	X	0.49	0.62
	EMBSD-L1	X	X	X	X	X	0.40	0.41
Tandem	EMBSD	X	0.24	0.51	X	0.38	X	X
	EMBSD-L1	X	0.25	0.55	X	0.39	X	X
Low-rate	EMBSD	X	X	X	X	X	0.38	0.35
	EMBSD-L1	X	X	X	X	X	0.50	0.49

Table 9 RMSE data for EMBSD and EMBSD-L1

By using POB averaging, EMBSD produces lower RMSE results for the error category. EMBSD-L1 predicts the no error category as well as, and in some cases, better than EMBSD. EMBSD-L1 performs better under MNRU conditions in four of seven databases. It might be due to the type of noise used in MNRU. The noise in MNRU is uncorrelated with the signal and is stationary throughout the signal. While EMBSD may perform better under bursty error than EMBSD-L1, EMBSD-L1 may perform better under uncorrelated, stationary noise.

Background noise can be stationary or bursty. The jeep noise in T7 and T8 can be categorized more as a consistent disturbance throughout the speech than a bursty type of distortion. The noise conditions have low RMSE when averaged by the EMBSD-L1. The tandem condition may be viewed as a special case of a background noise condition.

Distortions introduced by the first coder in the tandem play the role of added noise to the second coder.

For low rate coders, POB shows improvements over the L1 norm. In fact, there is a trend of lower RMSE for POB under categories where low-rate coders are prominent. In T7 and T8, containing a number of low rate coders, POB produces lower RMSE for no error, error, and low-rate coders. Even though L1 norm performs just as well as POB in a number of conditions, POB will be used in the remainder of this research because of the clear improvement it produces in the correlation analysis.

4.4 Predicting Quality Under Various Conditions for a Given Coder

The ability to predict speech quality under various conditions is important for an objective measure. Various distortion conditions were evaluated for a given coder. As only a small number of conditions are available in the databases for a given coder, the results may not be conclusive. However, the results lead to some interesting observations.

Figures 4(a) - (d) represent a sampling of the results obtained from the test databases used to evaluate EMBSD. Each figure shows both EMBSD and MOS as a function of the test condition.

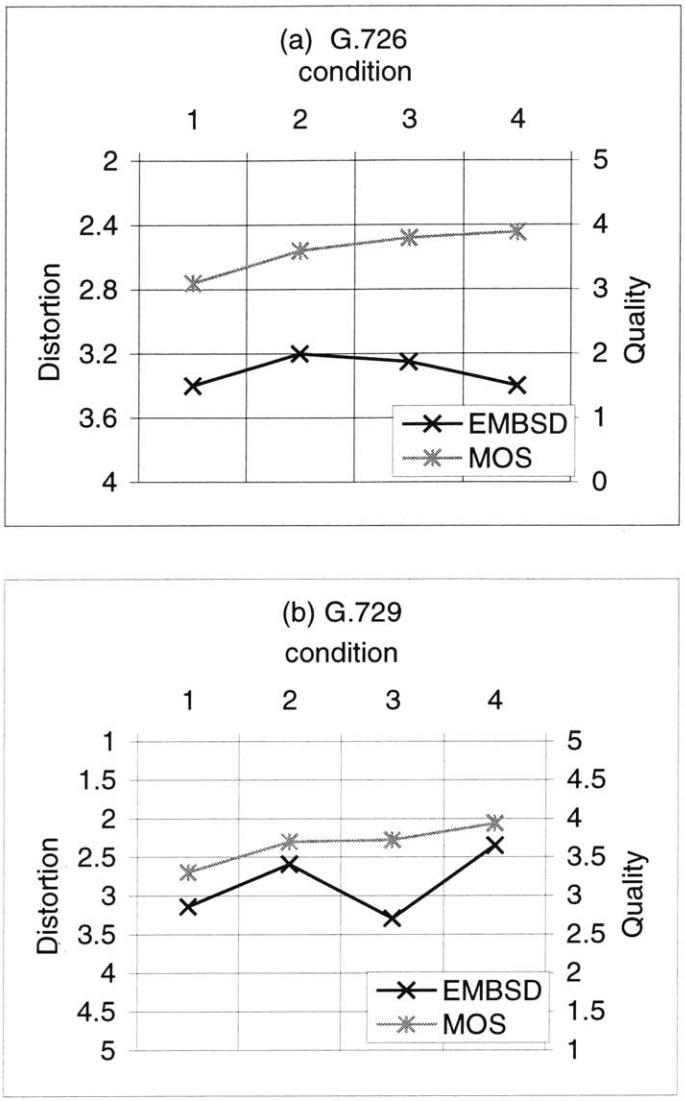

Figure 4 (a) and (b) Plots of EMBSD and MOS for Various Conditions

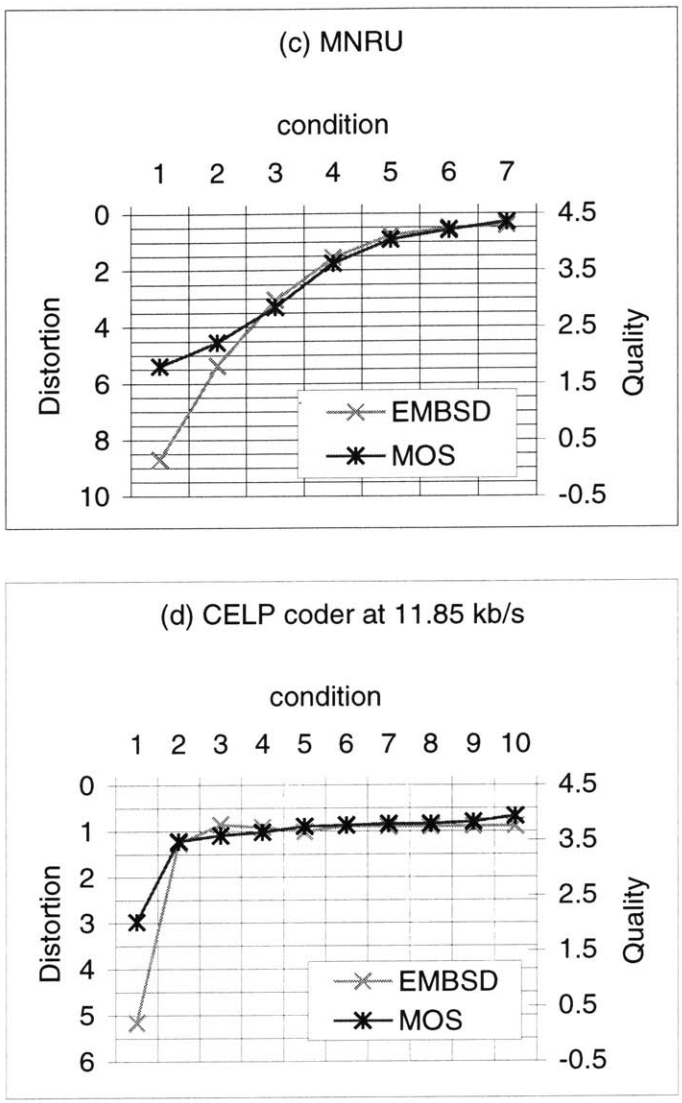

Figure 4 (c) and (d) Plots of EMBSD and MOS for Various Conditions

In each figure, conditions are ordered from the lowest to the highest MOS. The conditions include bit error, frame erasure, tandem, level variation, and MNRU. The right ordinate is a quality scale for MOS and the left ordinate is a distortion scale for EMBSD scores. The distortion axis is inverted so that the EMBSD and MOS curves slope in the same direction.

EMBSD scores seem to generally predict qualitative trend of various conditions. As shown in figures 4(a)-(d), EMBSD scores, in general, decrease in distortion when MOS increase in quality. Figure 4(a) and (b) shows that EMBSD has difficulty evaluating level variations, which are conditions 2, 3, and 4. A closer examination shows that the separation between the lowest and highest point is very small, only about 0.3 MOS. EMBSD is not able to predict such a small difference in this case.

4.5 Predicting Quality Across Coders

In standards competitions, subjective tests are used to choose the highest quality coder. If objective tests are to replace or augment subjective tests, it is important that they can rank coders accurately.

Figures 5(a)-(d) each show 4 coders: GSM Enhanced Full Rate, GSM Full Rate, 11.85 kb/s CELP, and 7.45 kb/s CELP.

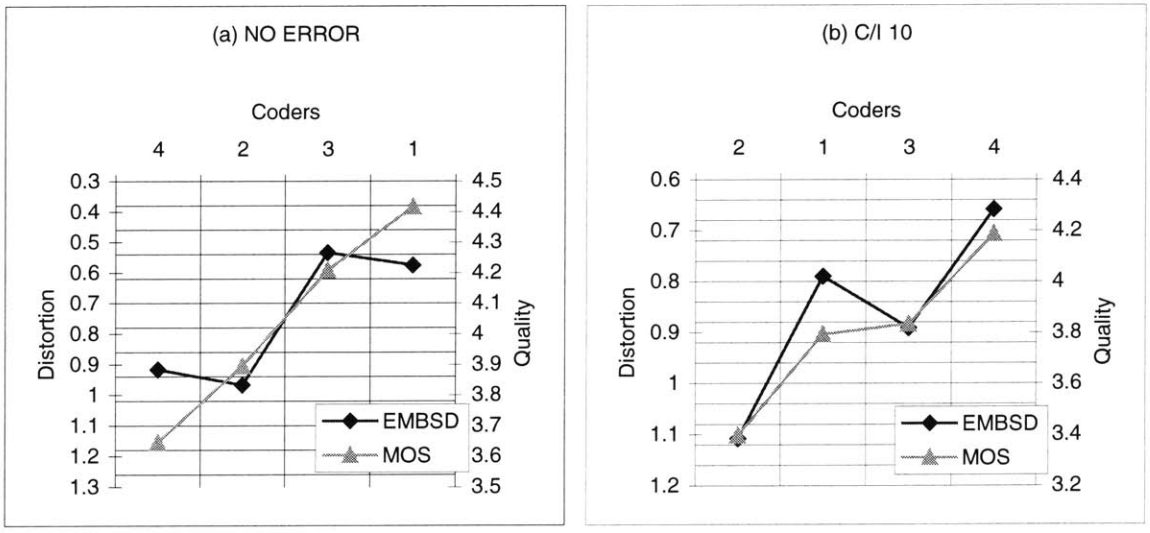

Figure 5 (a)-(b) Plots of EMBSD and MOS Across Coders

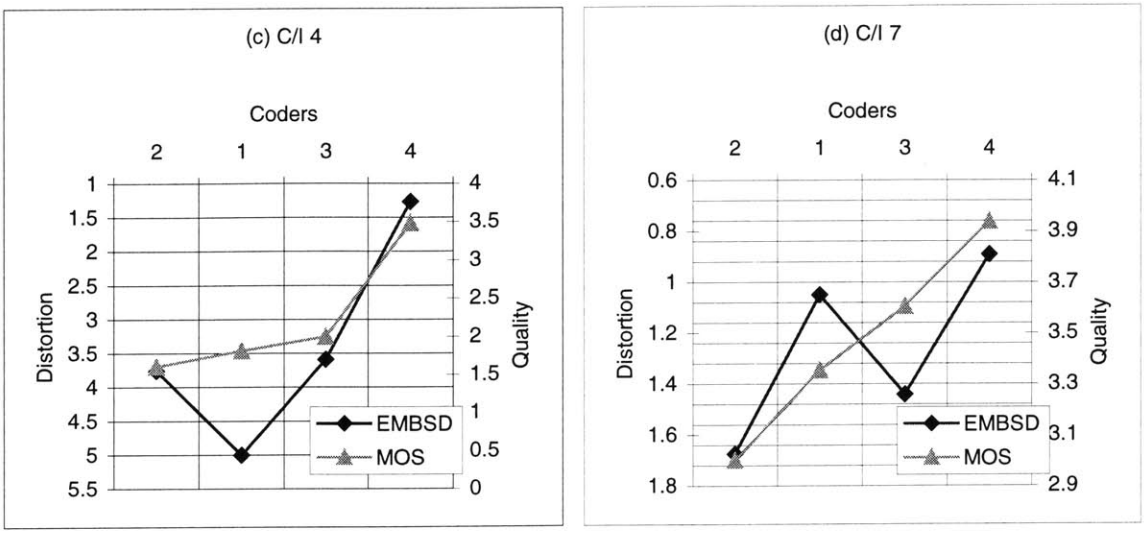

Figure 5 (c)-(d) Plots of EMBSD and MOS Across Coders

In each figure, the speech is tested under a different condition; figure 5(a) shows the no error condition and figures 5(b)-(d) show increasing bit error conditions.

The distortion axis used for EMBSD is on the left of the graph and the quality axis for MOS appears on the right. Grid lines are all 0.2 points apart based on the Quality scale and attempts were made to keep the distortion and quality axis on similar numerical scales to enable comparisons between figures. Coders were ordered in terms of increasing MOS. Coders 2 and 3 are properly aligned in all figures. Coders 2 and 4 are correctly aligned as well except in figure 5(a). However, incorrect orderings are also frequent, including coder 2 versus coder 4 in figure 5(a), coder 1 versus coder 3 in figures 5(a), 5(b), and 5(d), and coder 1 versus coder 2 in figure 5(c).

From the results presented in this and the previous section, it may be concluded that (1) EMBSD performs well in predicting speech quality of a given coder under different conditions and (2) EMBSD is unable to provide a consistently good prediction across different coders.

Chapter 5

EVALUATION OF PESQ

Chapter 5 focuses on the performance of PESQ, including the effectiveness of its time alignment mechanism.

5.1 Performance of PESQ

PESQ has been developed and optimized extensively for MOS prediction. PESQ was tested on the same eight databases used for testing EMBSD. The databases are summarized in Table 5, Section 4.3. Table 10 summarizes the results of correlation coefficients and RMSE between PESQ and MOS. PESQ scores are plotted against MOS in Figures 6(a)-(o). The results include PESQ and polynomial-mapped PESQ. In addition, a comparison between the correlation coefficients of PESQ and EMBSD is presented in Figure 7(a). The similar comparison for RMSE is shown in Figure 7(b). Results shown in Figures 7(a) and 7(b) refer to polynomial-mapped PESQ and EMBSD.

Database	PESQ		Polynomial-Mapped PESQ	
	Correlation	RMSE	Correlation	RMSE
T1	0.98	0.27	0.99	0.14
T2	0.92	0.26	0.96	0.16
T3	0.80	0.53	0.87	0.32
T4	0.88	0.40	0.91	0.26
T5	0.84	0.62	0.89	0.30
T6	0.84	0.45	0.85	0.39
T7	0.87	1.06	0.90	0.36
T8	0.87	0.90	0.89	0.37

Table 10 Correlation and RMSE data for PESQ before and after polynomial-mapping

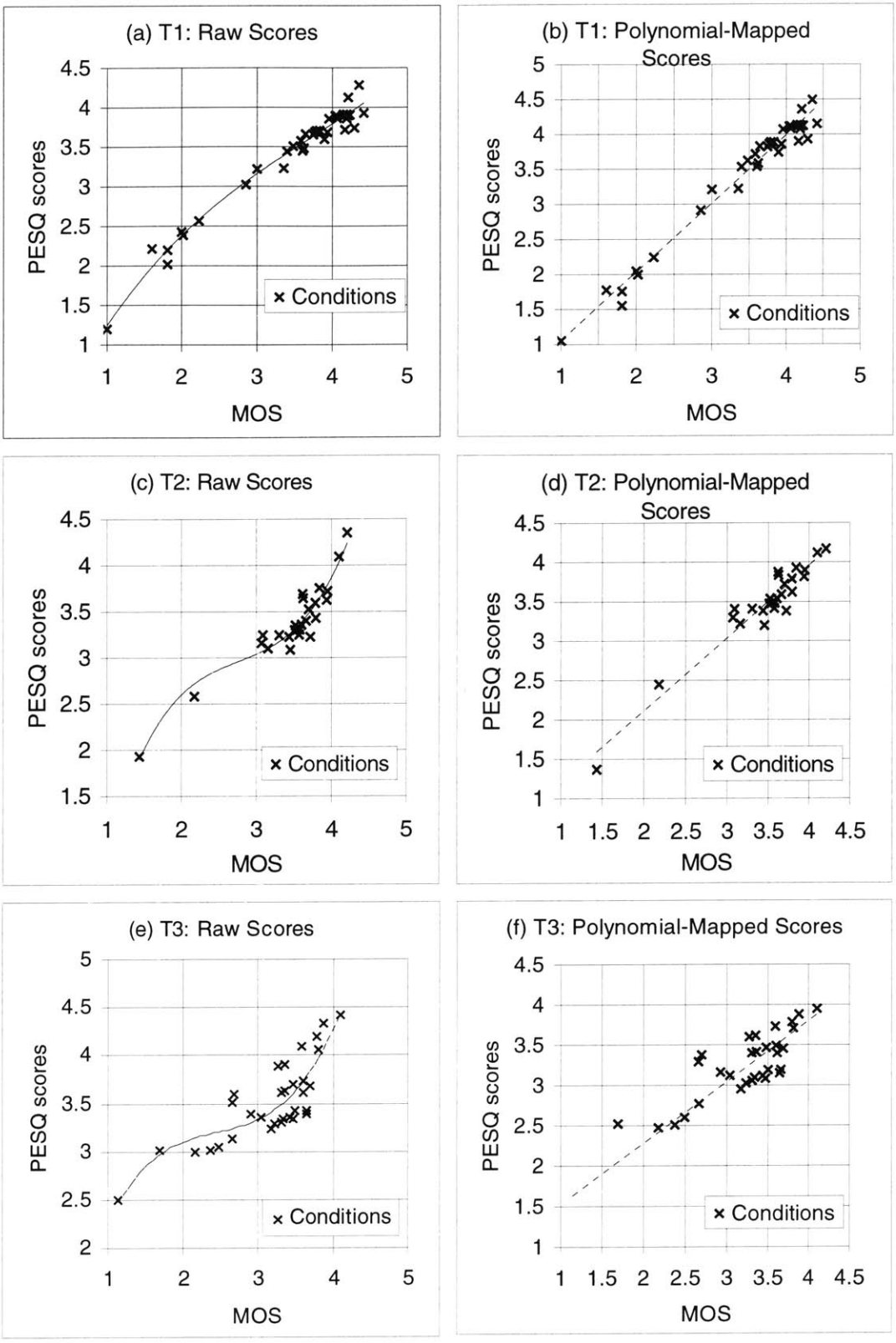

Figure 6 (a)-(f) Plots of PESQ vs. MOS before and after polynomial-mapping

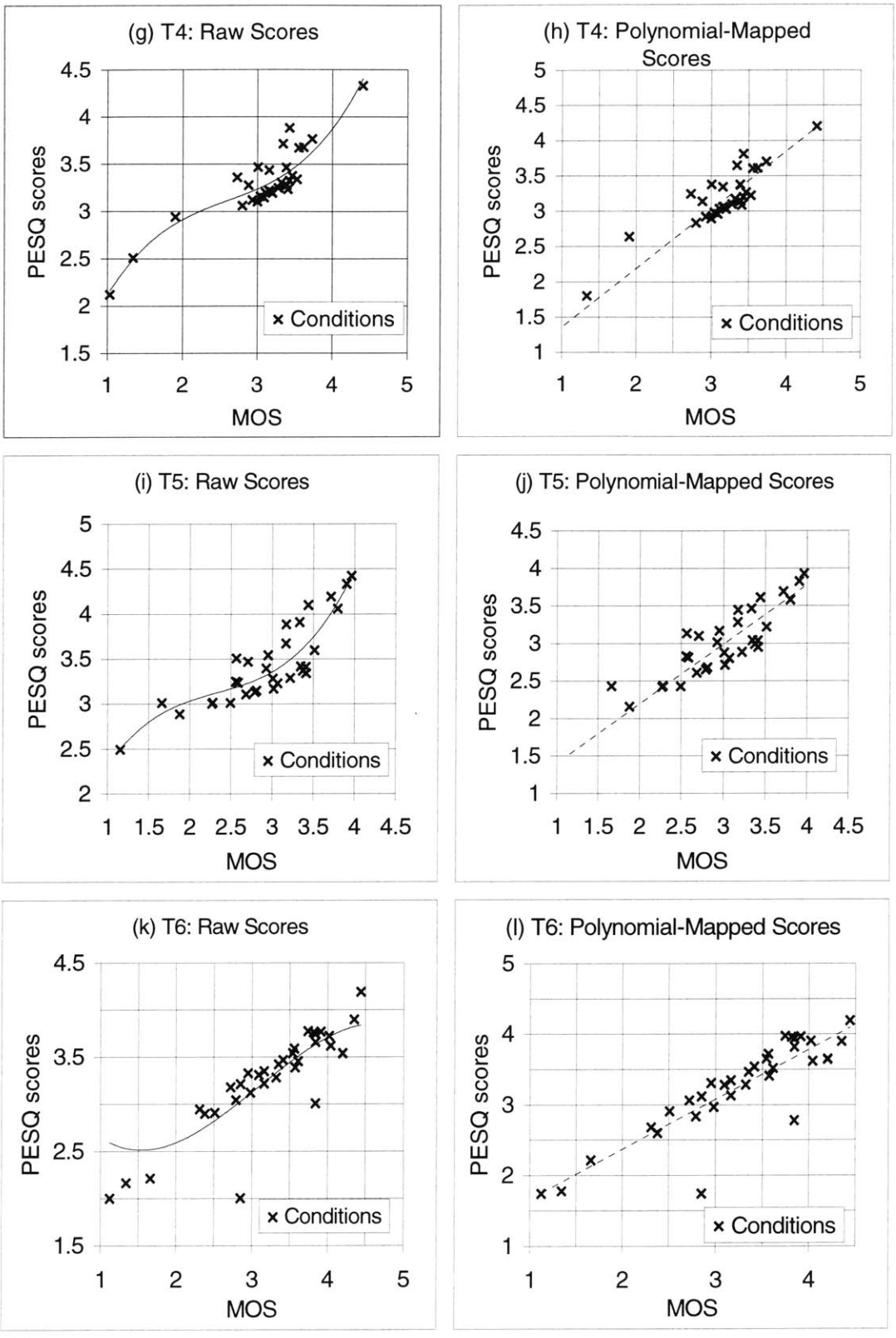

Figure 6 (g)-(l) Plots of PESQ vs MOS before and after polynomial-mapping

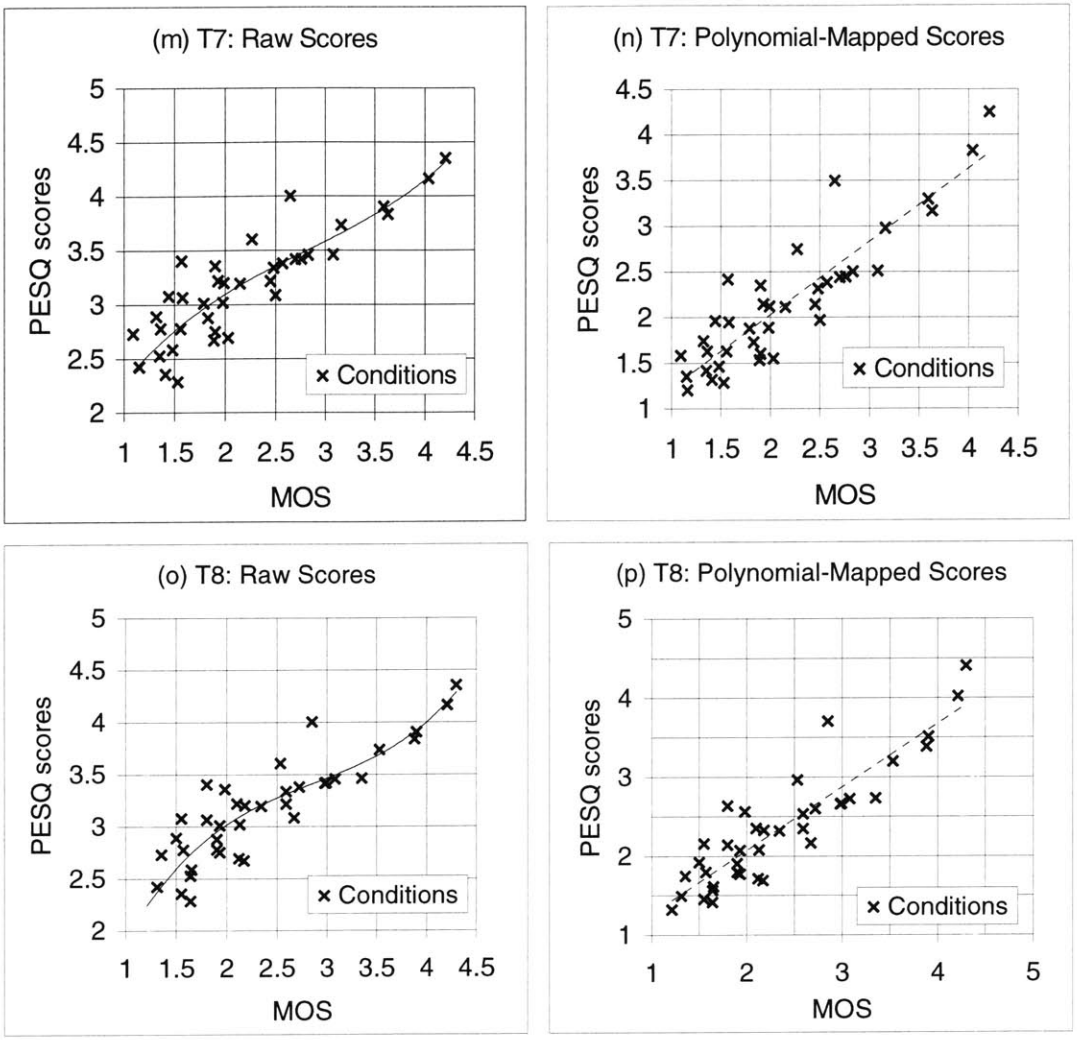

Figure 6 (m)-(p) Plots of PESQ vs. MOS before and after polynomial-mapping

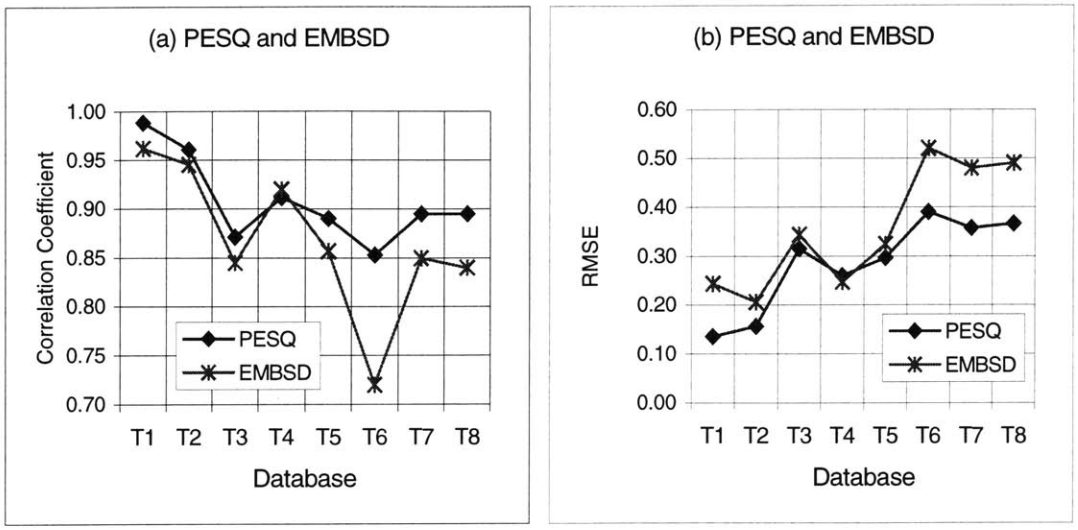

Figure 7(a)-(b) Correlation Coefficient and RMSE data for PESQ and EMBSD

Across the databases, PESQ shows a more consistently high correlation with MOS than EMBSD. RMSE results show that PESQ can predict MOS quite well in databases with good quality, such as T1, T2, and T4. In addition, PESQ provides higher correlations and lower RMSE than EMBSD in almost all cases, especially, in the databases with lower bit rate coders and poorer quality conditions, such as T6-T8.

5.2 Effectiveness of Time Alignment

In VoIP transmissions, there is a buffer at the receiver to hold incoming packets. The buffer reduces the effect of packet loss but also introduces delay. Practical systems try to balance packet loss and delay with dynamic algorithms that resize the buffer as needed. Buffer resizing changes the overall packet delay and causes a delay variation in packet flow. As a result, the transmitted signal is not aligned with the reference. If delay variations are too short to be perceived by the human ear, it is important that the

distortion is ignored by the objective measure. Many objective measures have difficulty distinguishing audible distortions from inaudible misalignments.

PESQ claims to be capable of handling variable delays in both silent and speech periods with a time alignment mechanism [13]. To evaluate the effectiveness of the time alignment mechanism, PESQ was evaluated under delay variation conditions in both silent and speech periods.

5.2.1 Delay Variations in Silent Periods

Data A has no delay variations and contains seven sentence pairs. Each sentence pair is encoded and decoded by the G.729 coder. To create Data B, 10 ms of delay or additional silence is inserted between encoding and decoding approximately every 160 ms only during a silent period. Since the human ear can tolerate up to 250 ms of delay before perceiving a drop in quality, A and B should sound identical [19]. If PESQ effectively aligns delay variations in silence periods, PESQ scores should be very similar.

	MOS Prediction	
	Database	
	A	B
PESQ	3.809	3.762
Difference from A	0	0.047

Table 11 Comparison between Databases A and B.

PESQ scores were averaged for both data sets and recorded in Table 11. Results show that the scores for A and B are practically the same. Thus, PESQ is able to align effectively speech with delay variations in silence periods.

5.2.2 Delay Variations in Speech Periods

Using A, Data C is created with delay variations in silence as well as speech periods. To process C, 10 ms of delay or silence is inserted every 160 ms between encoding and decoding. C is different from B because delay insertions occur in speech periods as well as during silence.

Delay variations in speech periods should make C sound degraded compared with A. However, PESQ scores of C cannot be evaluated directly because MOS for C is not available. An alternative approach was taken by creating Data D. D is the same as C except that a modified version of the G.729 decoder was used. The modified G.729 utilizes a different playback method, which treats delay variations differently from the G.729 used in C. The PESQ scores of C and D were then compared with subjective scores obtained with an A/B comparison test, in an effort to evaluate the capability of time alignment of PESQ. The subjective A/B Comparison test was run and the results are presented in Table 12. [1] PESQ scores for D and C are summarized in Table 13.

	C				D
	AB Comparison Preference Scale				
	Strong	Slight	No	Slight	Strong
Number of Votes	0	2	8	29	17

Table 12 A/B Test Results for Databases C and D

	Database	
	C	D
PESQ	2.830	2.795

Table 13 PESQ Scores for Databases C and D

[1] The subjective A/B Comparison test is similar to the CCR test. Listeners vote on a 5-category scale shown in Table 12. The A/B Comparison test is suitable for evaluating distortions between two sets of data.

AB Comparison test scores show that D is preferred to C by 80% and strongly preferred by 30%. In other words, there was a clear preference for D over C in this A/B listening test. If the time alignment mechanism of PESQ evaluates the different playback methods properly, PESQ scores of C and D should be consistent with that of the A/B comparison test, with D having a higher PESQ score than C. However, PESQ scores of D and C are very close. Further investigation is recommended to evaluate the PESQ time alignment mechanism, especially for delay variations during speech.

5.3 Predicting Quality Under Various Conditions For a Given Coder

As noted before, the ability to predict quality among various conditions is an integral part of objective measure performance. Similar to comparisons made in section 4.2.4, PESQ scores were compared to MOS under various conditions for a given coder. Polynomial-mapped scores were used in all cases. Figures 8 (a)-(d) show representative results obtained in this evaluation. In each figure, the conditions are ordered from lowest to highest MOS. The conditions include bit error, frame erasure, tandem, level variation, and MNRU.

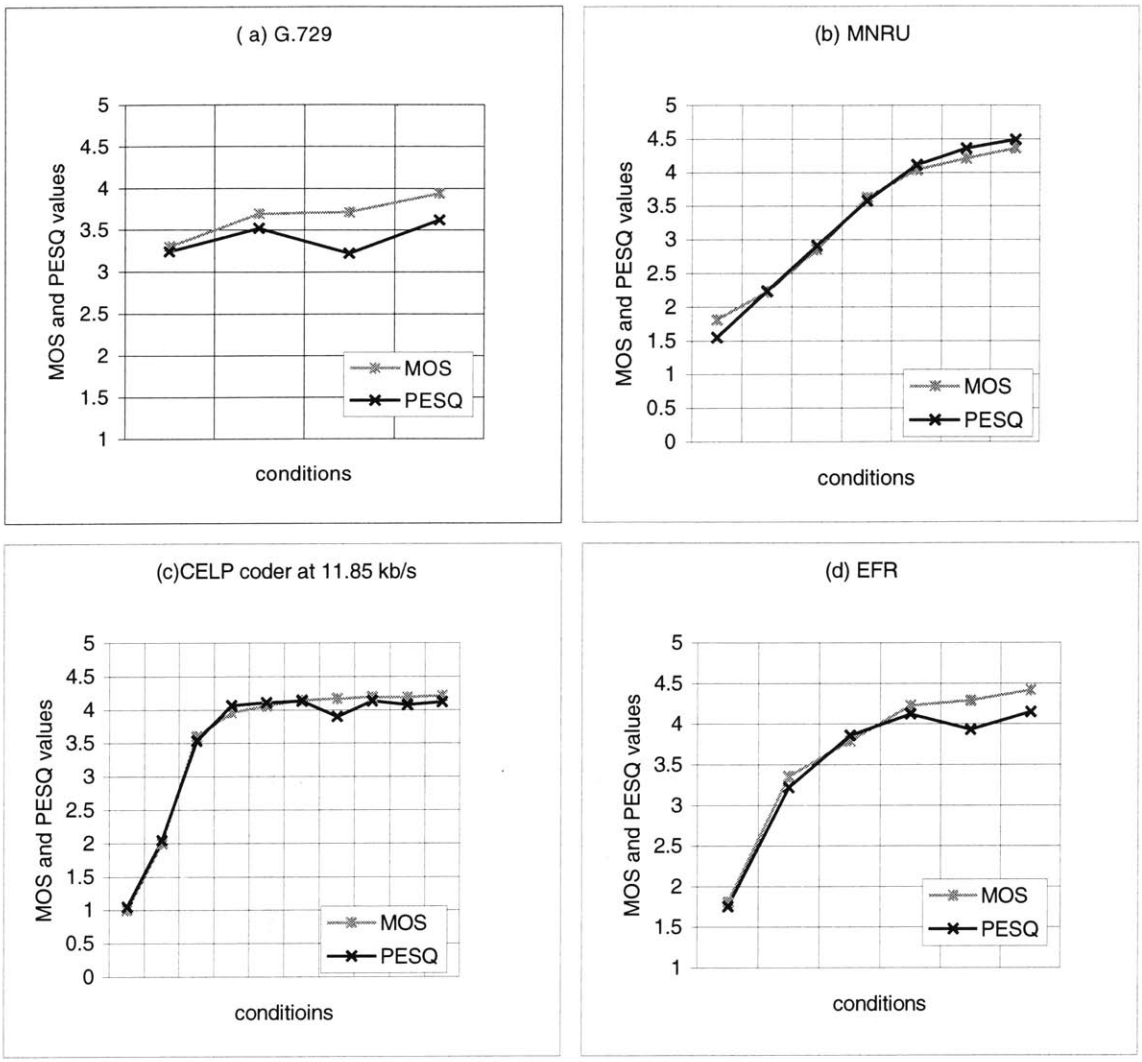

Figure 8(a) and (d) Plots of MOS and PESQ for Various Conditions

Figures 8(a)-(d) show that PESQ consistently orders the conditions correctly with MOS.

In figures 8(b)-(d), the scores are nearly the same. The accuracy of PESQ scores in figure 8(b) for MNRU suggests that PESQ can predict quite well the quality for conditions where the distortion is uncorrelated with speech. Results in figures 8(c) and (d) are aligned well overall, except condition seven in figure 8(c) and condition five in figure 8(d). Both conditions are low input level conditions. This suggests that PESQ may have problems handling to low input level conditions.

5.4 Predicting Quality Across Coders

If objective tests are to replace or augment subjective tests, it is important that they evaluate and rank coders accurately. The same coders from the previous section are used in the following figures. Figures 9(a)-(d) show results for the 4 coders.

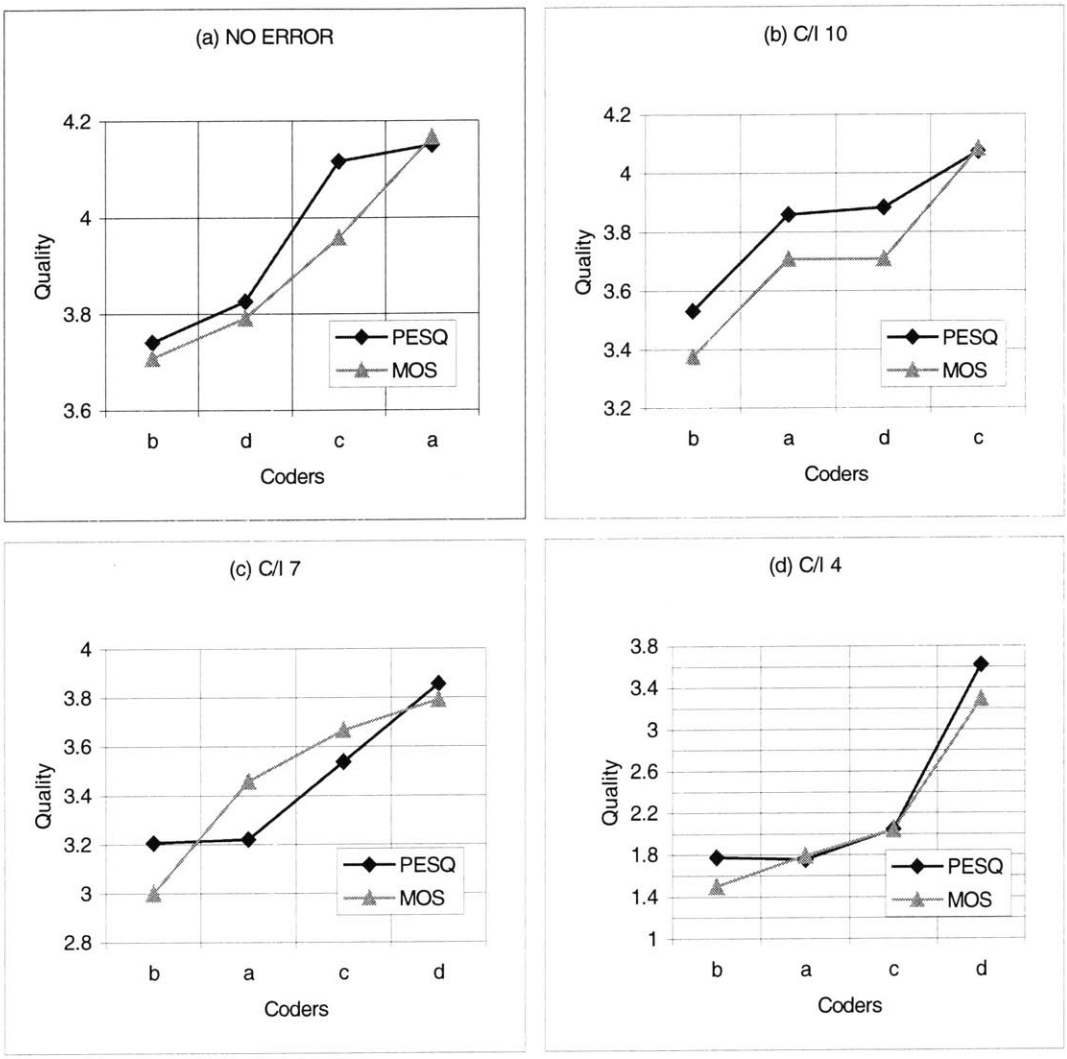

Figure 9(a)-(d) Plots of PESQ and MOS Across Coders

In each figure, the speech was tested under a different condition; figure 9(a) shows the no error condition and figures 9(b)-(d) show increasing bit error conditions.

Coders are ordered in terms of increasing MOS. In general, PESQ orders coders correctly with MOS. In figures 9(c) and (d), MOS for coder b is significantly lower than for coder a; however, PESQ is unable to capture those differences. Even though this points out that PESQ is not perfect, PESQ is much more effective in evaluating coders than EMBSD.

Chapter 6
WIDEBAND EXTENSION OF THE PESQ MEASURE

The results shown in the previous chapter suggest that PESQ is an effective objective measure. However, PESQ is recommended for the evaluation of narrowband speech, which is restricted to the 300 to 3400 Hz telephone bandwidth.

The desire for better than toll-quality speech has led to increasing demand for wideband speech. Wideband greatly improves quality by extending the bandwidth and dynamic range of narrowband speech. Wideband speech spans roughly 50 to 7000 Hz. The upper band extension gives a crisper and more intelligible speech while the lower band extension produces a more natural sound. Applications of wideband speech include audio and video teleconferencing, digital radio broadcasting, third generation wireless communications, and Voice over Internet Protocol (VoIP).

In 1988, the CCITT, which is now known as the ITU, established an international standard for high quality 7 kHz audio coding, known as G.722 [20]. Currently, the ITU is pursuing a new standard for wideband speech coding at 16 kb/s. The developments in wideband speech coding motivate the extension of the PESQ measure to handle wideband speech. This wideband extension is denoted here as PESQ-WB.

6.1 Extension of the PESQ Measure

From a thorough examination, it was determined that only a few changes were required

to extend PESQ to evaluate wideband speech. The changes included removing the telephone band filters and ensuring that the psychoacoustic mapping of spectra ranged from 50 to 7000 Hz.

6.2 Performance of PESQ-WB

Three wideband tests, T9-T11, were used to evaluate PESQ-WB. T9 includes error free, bit error, frame erasure, tandem, and various input level conditions. There are thirty-six conditions in total including six wideband MNRU conditions. The coders tested are versions of G.722 and a CELP-based wideband coder in various bit-rates ranging from 12 kb/s to 64 kb/s. Test T10 is similar to T9 except that the G.722 and wideband CELP coders range from 16 kb/s to 64 kb/s. There are a total of 24 conditions, 4 of which are wideband MNRU conditions. Test T11 evaluates various levels of static C/I conditions at the nominal input level. A wideband CELP coder and G.722 are tested at various bit rates along with 6 MNRU conditions.

Figures 10(a)-(f) show the results for Tests T9, T10, and T11. Figures 10(a), (c), and (e) show scatter plots between the MOS and PESQ scores, with the smooth curve showing the 3rd-order polynomial mapping.

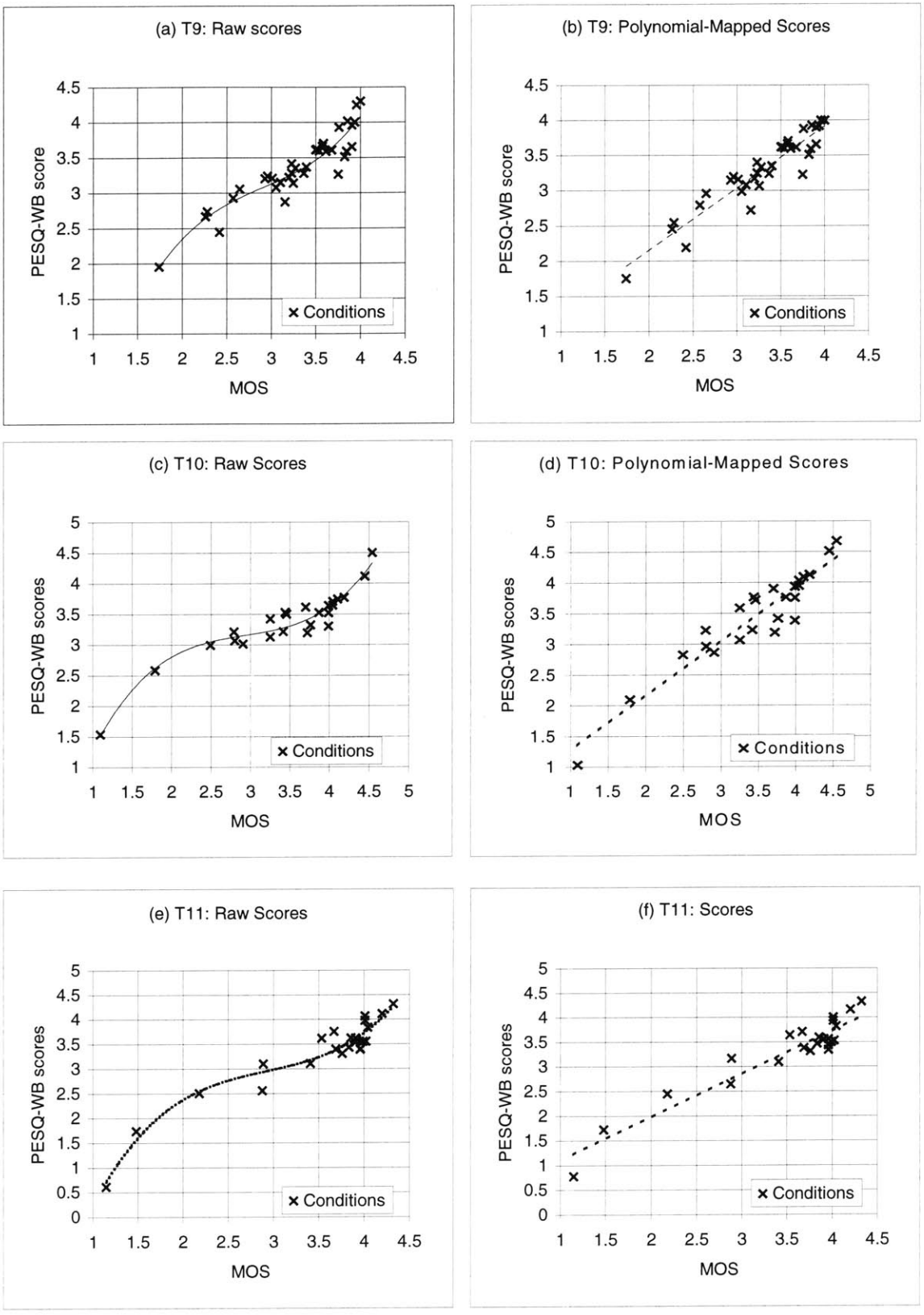

Figure 10 (a)-(f) Plots of PESQ-WB vs. MOS before and after polynomial-mapping

The performance of PESQ-WB was evaluated with correlation values and RMSE. Results are summarized in Table 14.

Database	PESQ-WB		Polynomial-mapped PESQ-WB	
	Correlation	RMSE	Correlation	RMSE
T9	0.922	0.224	0.938	0.189
T10	0.922	0.384	0.944	0.265
T11	0.946	0.337	0.965	0.213

Table 14 Correlation and RMSE data for PESQ-WB

The correlation values of PESQ-WB for wideband speech are higher than most of the correlation values of PESQ with narrowband speech shown in Table 10. The results demonstrate that PESQ-WB has a good potential for being an effective predictor of speech quality for wideband speech. Further investigation involving different wideband databases is recommended.

Chapter 7
DMOS PREDICTION

The goal of EMBSD, PESQ, and other objective measures is, in general, to predict MOS. However, the prediction of other subjective scores, such as DMOS, should also be considered. This section focuses on how well EMBSD and PESQ are able to predict DMOS.

DCR measures are typically used to evaluate speech under background noise conditions. There are three different input signals that can be used by the objective measure: the clean signal, the direct signal, and the processed signal as shown in Figure 11.

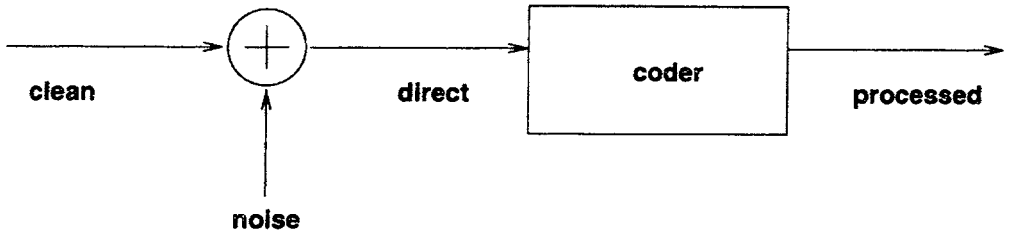

Figure 11 Speech under Background Noise Conditions

The clean signal is the clean source signal that is typically used as the reference in the objective measures to predict MOS. The direct signal is the clean signal with added noise. The processed signal is the direct signal after it has been processed by the coder.

In DCR tests, listeners rate the processed signal compared to the direct signal. Therefore, it is logical to use the direct and processed signals as the reference and distorted inputs, respectively, for an objective measure.

7.1 Combination Score

The presence of added noise, however, may affect the listener's perception of the speech quality. The listener may give lower ratings in DCR tests simply because the speech is noisy. To capture such an effect, the direct signals can be compared against clean signals in objective tests. Let (C,D) denote the objective test score that uses the clean and direct signals as reference and distorted inputs, respectively. Let (D,P) denote the score where direct and processed signals are used instead.

A combination of the foregoing two objective scores may predict perceived quality better than (D,P) score alone. The calculation of Combination score is shown in Equation 16.

$$Combination\ score = a[D,P] + (1-a)[C,D] \quad a \leq 1 \tag{16}$$

However, the (C,D) scores are often the same across all coder conditions since the same uncoded sentences, clean and direct signals, are used for each noise condition. The outcome makes the (C,D) term be a constant. Therefore, the correlation results for Combination scores will not differ from the correlation results for (D,P) scores when considering a single noise type. Let (C,P) denote the objective test score that uses the clean and processed signals as reference and distorted inputs, respectively. (Recall that PESQ recommends using (C,P) for MOS prediction.) Because the processed signal is coded, the (C,P) term will not be uniform across coder conditions.

$$Combination\ score = b[D,P] + (1-b)[C,P] \quad b \leq 1 \tag{17}$$

Below, only the Combination Score given in Equation 17 is used.

7.2 Performance of EMBSD

The performance of the EMBSD measure in DMOS prediction was evaluated with three databases, T12-T14. Both correlation and RMSE of the EMBSD scores were evaluated. T12 and T13 deal with street and car noise conditions, respectively. In both tests, the error conditions include various levels of static C/I conditions at a nominal input level. Both tests use five different coders within a range of 8 kb/s to 20 kb/s. The coders include G.729, GSM Full Rate, GSM Enhanced Full Rate, CELP coder at 5.15 kb/s, 7.45 kb/s, and 11.85 kb/s. Six MNRU conditions are also included. T14 deals with car noise conditions. All conditions were tested under error free conditions at a nominal input level. The test has 8 different coders ranging from 5.6 kb/s to 11.9 kb/s. The coders include CELP coders at 11.9 kb/s and 9.5 kb/s, Variable-rate CELP at 9.6 kb/s and 5.8 kb/s, PCS1900, G.728, GSM Full Rate, and GSM Half Rate. Two MNRU conditions are also included in the test. The results are summarized in Table 15.

Database	Objective Score	EMBSD		Polynomial-Mapped EMBSD	
		Correlation	RMSE	Correlation	RMSE
T12	(D,P)	0.750	0.798	0.836	0.456
	Combination	0.772	2.174	0.830	0.491
	(C,P)	0.770	2.549	0.849	0.441
T13	(D,P)	0.799	0.832	0.850	0.492
	Combination	0.816	1.079	0.862	0.474
	(C,P)	0.814	2.230	0.896	0.414
T14	(D,P)	0.740	0.556	0.776	0.440
	Combination	0.811	0.535	0.813	0.421
	(C,P)	0.827	0.501	0.893	0.301

Table 15 DMOS Prediction data for EMBSD (b=0.30)

The results show that (C,P) scores correlate better with DMOS than do the Combination and (D,P) scores. Combination scores correlate with DMOS better than (D,P) scores alone. The (C,P) measure also produces, in general, low RMSE values, for polynomial-mapped EMBSD.

7.3 Performance of PESQ

The performance of the PESQ measure in the prediction of DMOS used the same three test databases as in the evaluation of EMBSD. In addition, PESQ-WB is evaluated by three other wideband DCR databases, T15, T16, and T17.

7.3.1 Narrowband Speech Data

Database	Objective Score	PESQ		Polynomial-Mapped PESQ	
		Correlation	RMSE	Correlation	RMSE
T12	(D,P)	0.905	0.749	0.943	0.277
	Combination	0.920	0.943	0.929	0.308
	(C,P)	0.720	1.470	0.850	0.438
T13	(D,P)	0.917	0.790	0.961	0.256
	Combination	0.926	0.912	0.970	0.225
	(C,P)	0.844	1.246	0.944	0.308
T14	(D,P)	0.801	0.444	0.832	0.396
	Combination	0.850	0.841	0.949	0.226
	(C,P)	0.677	1.142	0.678	0.525

Table 16 DMOS Prediction data for PESQ (b=0.70)

Unlike the outcome for the EMBSD measure, the PESQ measure clearly shows that Combination scores correlate the best with DMOS. The percentage of the (D,P) and (C,P) scores leading the best correlation differs among the three tests; however, a 70/30 percentage split favoring the (D,P) score produces good overall results for all tests.

The RMSE for Combination scores are larger than for (D,P) before the polynomial mapping function was applied. However, results after cubic mapping show the RMSE for the Combination scores are lower than for (D,P) and for (C,P) for Test 2 and Test 3. Results in Table 16 also show that (D,P) scores are consistently more effective than (C,P) scores.

7.3.2 Wideband Speech Data

Three wideband tests, T15-T17, were used to evaluate the performance of PESQ-WB on DMOS prediction with many different noise conditions. T15 and T16 evaluated street and car noise conditions, respectively. In both tests, the error conditions included various levels of static C/I conditions at a nominal input level. Six coders were tested. They were three wideband coders under test and the G.722 at 64, 56, and 48 kb/s. MNRU conditions were also part of the two tests. T17 evaluated four different types of noise: office, babble, car, and interference talker. Error conditions were not included in the test. Five coders were tested with bit rates ranging from 12 kb/s to 56 kb/s. The coders included G.722 at 48 and 56 kb/s, and a CELP-based wideband coder.

Database	Objective Score	PESQ-WB		Polynomial-Mapped PESQ-WB	
		Correlation	RMSE	Correlation	RMSE
T15	(D,P)	0.901	0.444	0.926	0.348
	Combination	0.810	0.950	0.852	0.482
	(C,P)	0.367	1.142	0.511	0.790
T16	(D,P)	0.884	0.715	0.929	0.318
	Combination	0.832	1.024	0.884	0.402
	(C,P)	0.453	1.889	0.619	0.673
T17	(D,P)	0.886	0.394	0.914	0.279
	Combination	0.904	0.636	0.921	0.268
	(C,P)	0.391	1.497	0.515	0.590

Table 17 DMOS Prediction data for PESQ-WB (b= 0.70)

Unlike PESQ narrowband results, Combination scores do not perform as well as (D,P) scores. For the tests T15 and T16, Combination scores are significantly worse than (D,P) scores. For test T17, Combination scores show some improvement, but the improvement is not as great as in the narrowband tests. However, like PESQ, PESQ-WB produces consistently higher correlations with (D,P) scores than with (C,P) scores. As shown in Table 17, the difference between correlation coefficients of the two cases is very dramatic. On average the (D,P) scores correlate at 0.89 while the (C,P) correlate at 0.40 only.

Chapter 8

CONCLUSION

In this thesis, we evaluated the EMBSD and PESQ objective speech quality measures. In the investigation of EMBSD, we have found that forward masking did not show satisfactory results. L1 averaging was found to be better at evaluating the no error condition and conditions where the distortion was stationary. The time alignment mechanism in PESQ is suitable for evaluating delay variations in silent periods; however, it does not seem as effective for evaluating delay variations in speech periods. PESQ-WB showed promising results in evaluating wideband speech. Future developments of wideband coders will find the results useful. In addition to predicting MOS, it was found that EMBSD and PESQ could predict DMOS.

Further research is warranted in the following areas: use of forward masking in wideband speech evaluation, improving EMBSD in evaluating different coders, and performing a more thorough testing of PESQ's time alignment mechanism using data with MOS scores.

9 REFERENCES

[1] Methods for subjective determination of transmission quality. ITU –T Recommendation P.800, August 1996.

[2] C. Jin, and R. Kubicheck. "Vector quantization techniques for output-based objective speech quality." IEEE ICASSP 1996.

[3] S.Quackenbush, T. Barnell III, and M. Clements. Objective measures of speech quality. Englewood Cliffs: Prentice Hall, 1988.

[4] J. Makhoul , R. Viswanathan, and W. Russell. "A framework for the objective evaluation of vocoder speech quality." IEEE ICASSP 1976.

[5] F. Itakura. "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech and Signal Processing, vol. ASSP-23, no. 1, pp. 67-72, February 1975.

[6] R. Viswanathan, W. Russell, and J. Makhoul, "Objective Speech Quality Evaluation of Narrowband LPC Vocoders," Proc. 1978 IEEE ICASSP, Tulsa, April 1978.

[7] T. Painter, A. Spanias. "Perceptual coding of digital audio." Proceedings of the IEEE, vol. 88, no.4, April 2000.

[8] S. Wang, A. Sekey and A. Gersho. "An objective measure for predicting subjective quality of speech coders." IEEE Journal on Selected Areas in Communications, Vol. 10, No. 5, pp. 819-829, June 1992.

[9] W. Yang. "Enhanced modified bark spectral distortion (EMBSD): an objective speech quality measure based on audible distortion and cognition model." Ph.D. Thesis. Temple University, May 1999.

[10] E. Zwicker and H. Fastl. Psychoacoustics: Facts and Models. Springer-Verlag: Berlin 1990.

[11] A.W. Rix and M.P. Hollier. "The perceptual analysis measurement system for robust end-to-end speech quality assessment." ICASSP 2000.

[12] J. Beerends and J. Stemerdink. "A perceptual speech-quality measure based on a psychoacoustic sound representation." J. Audio Engineering Society, vol. 42, no. 3, March 1994.

[13] A.W. Rix and M.P. Hollier. "PESQ – the new ITU standard for end-to-end speech quality assessment." To be presented at the 109th Convention, Los Angeles, US, 2000 September 22-25.

[14] Objective quality measurement of telephone-band (300-3400 Hz) speech codecs. ITU-T Recommendation P.861, February 1998.

[15] A.W. Rix, R.J. Reynolds, and M.P. Hollier. "Perceptual measurement of end-to-end speech quality over audio and packet-based networks." Presented at the 106th Convention Munich, Germany, May 1999.

[16] J. Beerends. "Audio Quality Determination Based on Perceptual Meausrement Mechniques." Applications of Digital Signal Processing to Audio and Acoutics. Edited by Mark Kahs and Karlheinz Brandenburg. Boston: Kluwer Academic Publishers, 1998.

[17] D. Kim; O. Ghitza; P. Kroon. "A Computational Model for MOS Prediction." 1999 IEEE Speech Coding Workshop.

[18] L. Thorpe and W. Yang. "Performance of Current Perceptual Objective Speech Quality Measures." 1999 IEEE Speech Coding Workshop.

[19] "Quality of Service Testing in the VoIP Environment." http://www.empirix.com/empirix/voice+network+test/resources/qos+testing+for+voip.html.

[20] Maitre. "7 kHz audio coding within 64kbit/s." IEEE Journal on Selected Areas in Comm., vol, 6, No. 2 pp. 283-298, February 1988.