3-27-2013 12:00 AM

Objective and Subjective Evaluation of Wideband Speech Quality

Nazanin Pourmand
The University of Western Ontario

Supervisor
Dr. Vijay Parsa
The University of Western Ontario

Graduate Program in Electrical and Computer Engineering
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Objective and Subjective Evaluation of Wideband Speech Quality

(Thesis Format: Monograph)

by

Nazanin Pourmand

Graduate Program in Engineering Science
Department of Electrical and Computer Engineering

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Traditional landline and cellular communications use a bandwidth of 300 - 3400 Hz for transmitting speech. This narrow bandwidth impacts quality, intelligibility and naturalness of transmitted speech. There is an impending change within the telecommunication industry towards using wider bandwidth speech, but the enlarged bandwidth also introduces a few challenges in speech processing. Echo and noise are two challenging issues in wideband telephony, due to increased perceptual sensitivity by users.

Subjective and/or objective measurements of speech quality are important in benchmarking speech processing algorithms and evaluating the effect of parameters like noise, echo, and delay in wideband telephony. Subjective measures include ratings of speech quality by listeners, whereas objective measures compute a metric based on the reference and degraded speech samples. While subjective quality ratings are the "gold-standard", they are also time- and resource- consuming. An objective metric that correlates highly with subjective data is attractive, as it can act as a substitute for subjective quality scores in gauging the performance of different algorithms and devices.

This thesis reports results from a series of experiments on subjective and objective speech quality evaluation for wideband telephony applications. First, a custom wideband noise reduction database was created that contained speech samples corrupted by different background noises at different signal to noise ratios (SNRs) and processed by six different noise reduction algorithms. Comprehensive subjective evaluation of this database revealed an interaction between the algorithm performance, noise type and SNR. Several auditory-based objective metrics such as the Loudness Pattern Distortion (LPD) measure based on the Moore - Glasberg auditory model were evaluated in predicting the subjective scores. In addition, the performance of Bayesian Multivariate Regression Splines(BMLS) was also evaluated in terms of mapping the scores calculated by the objective metrics to the true quality scores. The combination of LPD and BMLS resulted in high correlation with the subjective scores and was used as a substitution for fine-tuning the noise reduction algorithms.

Second, the effect of echo and delay on the wideband speech was evaluated in both listening and conversational context, through both subjective and objective measures. A database containing speech samples corrupted by echo with different delay and frequency response characteristics was created, and was later used to collect subjective quality ratings. The LPD - BMLS objective metric was then validated using the subjective scores.

Third, to evaluate the effect of echo and delay in conversational context, a realtime simulator was developed. Pairs of subjects conversed over the simulated system and rated the quality of their conversations which were degraded by different amount of echo and delay. The quality scores were analysed and LPD+BMLS combination was found to be effective in predicting subjective impressions of quality for condition-averaged data.

KEYWORDS: Speech quality evaluation, Wideband speech, Subjective test, Objective measures, Auditory model, Noise reduction algorithm, Real-time simulator.

Acknowledgement

First and foremost, I wish to express my sincere gratitude and appreciation to my supervisor, Dr. Vijay Parsa, for his continuous support, guidance and motivation throughout the research of this thesis. I would like to thank him for his attention and patience.

I am also grateful for the help and support from people in Balckberry. A special thanks to Chris Forrester and Leigh Thorpe for providing instructive comments and insightful feedback on my research. Sincere thanks also go to Sylvain Angrignon and Malay Gupta for their help.

I would also like to thank the members of the examining committee, Dr. Martin Bouchard, Dr. Susan Scollie, Dr. Hanif Ladak, and Dr. Serquei L. Primak for their time in reviewing this thesis and their comments. A special thanks to Dr. Duncan Murdoch for being in my advisory committee and his help and feedback on Bayesian analysis.

I am thankful to all the people in National Centre for Audiology, who made this environment a great place to work. Thanks to Steve Beaulac for his help and guidance on developing the software. My gratitude goes to my friend, Laya Poost-Foroosh Bataghva, whose experience helped me through these years.

I wish to thank my parents, who encouraged and supported me from far away. I wish I could show them how much I love and appreciate them.

Last but not the least, I would like to thank my husband, Iman khalaji, for his love and continuous support from day one.

Funding acknowledgment

I would like to thank Blackberry, Ontario Research Fund and Western Graduate Research Scholarship for providing financial support.

Table of Contents

List of Tables

List of Figures

Nomenclature

IS	Itakura saito
LPD	Loudness Pattern Distortion
CCITT	International Telegraph and Telephone Consultative Committee
ISDN	Integrated Services Digital Network
PSTN	Public Switched Telephone Network
ANSI	American National Standards Institute
BWE	Bandwidth Extension
ACR	Absolute Category Rating
MOS	Mean Opinion Score
HATS	Head and Torso Simulator
MUSHRA	MUltiple Stimulus test with Hidden Reference and Anchors
CQS	Continuous Quality Scale
PESQ	Perceptual Evaluation of Speech Quality
PESQM	Perceptual Evaluation of Speech Quality Measure
SNR	Signal to Noise Ratio
BMLS	Bayesian Multivariate Regression Splines
SNRseg	Segmental Signal-to-Noise Ratio
LPC	Linear Predictive Coefficients
LLR	Log-Likelihood Ratio
IS	Itakura Saito
SD	Spectral Distance
Log SD	Log Spectral Distance
WSS	Weighted Spectral Slopes
PEMO-Q	Perceptual Model – Quality Assessment
VAD	Voice Activity Detector
BM	Basilar Membrane

ERB	Equivalent Rectangular Bandwidth
PSM	Perceptual Quality Measure
HASQI	Hearing-Aid Speech Quality Index
CC	Cepstrum Correlation
MMSE	Minimum Mean Square Error
MARS	Multivariate Adaptive Regression Spline
MCMC	Markov Chain Monte Carlo
MB	Multiband Spectral Subtraction
SSNR	Segmental Signal to Noise Ratio
WSSD	Weighted Slope Spectral Distance
PLSR	Partial Least Squares Regression
DFT	Discrete Fourier Transform
KLT	Karhunen-Loeve Transform
LogMMSE	Log Minimum Mean Square Estimation
WCosh	Weighted Cosh
ICC	Intraclass Correlation Coefficient
ANOVA	ANalysis Of VAriance
FDR	False Discovery Rate
CASP	Computational Auditory Signal-processing and Perception
DRNL	Dual Resonance Nonlinear
MSPE	Mean Square Prediction Error
FFT	Fast Fourier Transform
SLR	Send Loudness Rating
RLR	Receive Loudness Rating
TELR	Talker Echo Loudness Rating
ERL	Echo Return Loss
RIM	Research In Motion, Inc.
AIR	Aachen Impulse Response database
RIR	Room Impulse Response
HHP	Hand-Held Position
HFRP	Hands-free Reference Point
EC	Echo Canceller
ES	Echo Suppressor
NLMS	Normalized Least Mean Square
FLMS	Fast Least Mean Square
MDF	Multidelay block Frequency
NLP	Nonlinear Processor
IRS	Intermediate Reference System
FIR	Finite Impulse Response
IPP	Integrated Performance Primitives
SPL	Sound Pressure Level
MRP	Mouth Reference Point
SRMR	Speech to Reverberation Modulation Energy Ratio

Chapter 1

Introduction

1.1 Introduction

Traditional landline and cellular communications use a bandwidth of 300–3400 Hz for transmitting speech [1]. This telephone band was determined by CCITT (International Telegraph and Telephone Consultative Committee) in the 1960s as a trade-off between technical limitations, transmission quality and economics [2]. Although adequate for speech communication, this narrow bandwidth of 300–3400 Hz has an impact on both the quality and intelligibility of transmitted speech. Wider bandwidth , 0 – 8 kHz , has been deployed within recently developed wireless and internet communication devices. The increase in speech communication bandwidth results in significantly better voice quality and eases the task of communication in noisy environments.

Most cellular and VoIP phone manufacturers that employ wideband telecommunications are interested in objective methods of quantifying the quality of their devices. The wideband speech quality assessment methods are also needed in order to compare and evaluate the performance of wideband noise reduction algorithms, wideband multi-microphone array processing algorithms, and wideband echo cancellation algorithms.

The purpose of this chapter is to: 1) review the advantages and challenges which accompany the bandwidth extension, 2) introduce speech quality evaluation and the available methods, 3) layout the scope of this thesis, and finally 4) outline the thesis organization.

1.2 Wideband Telephony

The so – called "wideband telephony" aims to extend the speech bandwidth to 50 - 7000 Hz, with a concomitant increase in quality and intelligibility. At the low end the speech is assumed to be uncontaminated by the power line interference, while the upper end is determined by the sampling theorem. In wideband telephony, the sampling rate is 16 kHz, according to the Nyquist sampling theorem, the upper limit could be as much as 8 kHz. The 50 – 7000 Hz passband has been specified in CCITT recommendation G. 722 [3].

End-to-end digital networks, such as the second and third generation wireless systems, Integrated Services Digital Network (ISDN), and voice over packet networks, do allow the use of wider speech bandwidth which results in communication quality significantly beyond that of the traditional Public Switched Telephone Network (PSTN) [4]. Wideband speech codecs have been standardized and are being used, providing significant improvements in terms of speech intelligibility and naturalness [5].

Wideband speech has already been deployed across enterprise networks using communication products from companies like Cisco and Avaya. It also has been used on the internet with PC-based VOIP phones (e.g. skype) [6] and there is a thrust in telecommunication industry towards using wider bandwidth speech in all transmission systems; but this upgrade in speech communication bandwidth also introduces some challenges in speech processing. In the following sections, the advantages and challenges presented by wideband speech are further explored.

1.2.1 Benefit

Wideband speech has lots of attractive features to offer. Better task performance and higher preference are two main advantages of wideband speech [6].

Better task performance - Wideband speech increases speech intelligibility, i.e. it improves the ability of understanding the meaning of a spoken message. Wideband speech makes this improvement in a couple of ways. First, wideband speech increases the human ability to separate the speech from the background noise. This process, which is called auditory streaming, is improved by having more information about the spatial properties of the speech and noise. Since wideband speech has more low and high frequency information it could be localized better than narrowband speech. Low frequency (< 1500 Hz) components provide interaural time difference cues and high frequency (> 1500 Hz) por-

tions deliver interaural level difference cues for sound localization, and these cues are better extracted from wideband speech than its narrowband counterpart [6][1].

Second, wideband speech provides more cues for recognizing phonemes, syllables and words within a stream. According to a Polycom report [7], "Two-thirds of the frequencies in which the human ear is more sensitive and 80 percent of the frequencies in which speech occurs are beyond the capabilities of the public telephone network. The human ear is most sensitive at 3.3 kHz, just where the telephone network cuts off". The energy in consonant sounds is primarily carried in the higher frequencies. Fricatives such as /s/, /sh/, /f/, whose spectral energy extends beyond 3400 Hz, are more affected by the limited bandwidth. The higher frequencies in wideband speech help discriminate these consonants. For example, the difference between /s/ and /f/ is detected in the frequencies above 3 kHz. Some commonly confused pairs are /p/ and /t/, /s/ and /f, /m/ and /n/ and so on [7]. Stelmachowicz *et al.* [8] showed that normal hearing adults achieved only 33% accuracy in identifying the /s/ phoneme spoken by a female talker when the bandwidth was 5 kHz and this improved to 80% with a bandwidth increase to 6 kHz.

The energy spectrum of a voiced and unvoiced speech is given in Figure 1.1 and Figure 1.2. Blue lines show the band limits of the narrowband telecommunication system and red lines demarcate the limits for wideband telephony. As can be seen in Figure 1.1, useful spectral content of the consonant /b/ exists at frequencies lower than 300 Hz. In addition, as shown in Figure 1.2, the consonant /s/ has significant frequency content beyond 3.4 kHz; all of this information will be filtered out by the narrowband system, thereby impacting their perception.

At a more global level, according to the American National Standards Institute (ANSI) standard for computing the speech intelligibility index [2], a further 18% increase in speech intelligibility for average speech is obtained when the bandwidth extends beyond 3400 Hz to 6500 Hz. It also has been shown that wideband speech facilitates speaker recognition [4] and improves speech recognition task [9, 10].

Higher preference- Wideband speech quality is usually rated higher in comparison with narrowband speech [11]. In addition to the increase in intelligibility, wider bandwidth is also associated with increased "brightness", "naturalness", and overall quality of speech [12, 13, 14]. For example, Moore and Tan [13] reported poor quality ratings for speech with narrow bandwidth and a 3-fold increase in perceived naturalness ratings of male and female

[1]It should be noted here that better auditory streaming and speech localization as the result of incorporating higher frequencies, only apply for binaural systems; while the focus of this thesis is on mono-channel wideband speech processing such as in basic telephony.

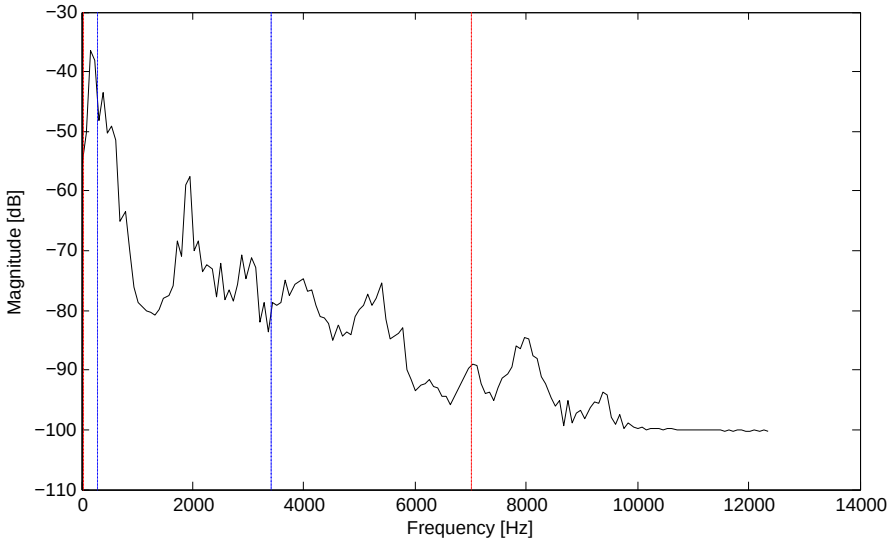

Figure 1.1: Energy spectrum of consonant /b/ in the word /ABIL/

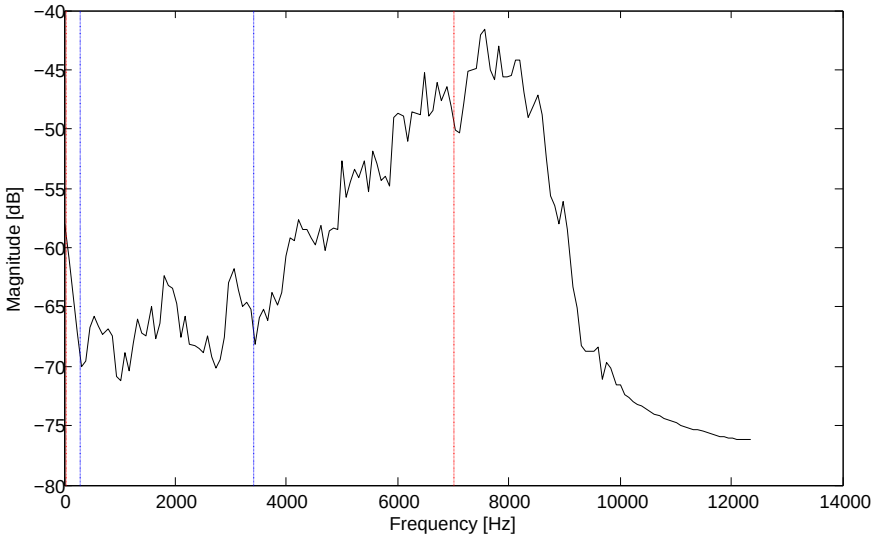

Figure 1.2: Energy spectrum of consonant /s/ in the word /ASIL/

speech samples when the bandwidth was increased from approximately 300–3400 Hz to 55–7000 Hz. In addition, according to AT&T [15] it is more pleasant and less fatiguing to listen to 7 kHz bandwidth speech in comparison with telephone bandwidth speech.

According to the AT&T technical journal [16], low-frequency extension (i.e. 50 to 300 Hz) contributes to improved naturalness and speaker recognition and high-frequency enhancement (i.e. 3400 to 7000 Hz) provides greater intelligibility and fricative discrimination (for example /s/ versus /f/).

The perceived speech quality of 19 speech pass-bands including narrowband and wideband

speech have been measured and compared in [2]. The comparison test showed that the bandwidth extension below 300 Hz improves the speech quality. It was also reported that extending the upper limit alone does not have significant effect on the perceived quality of the speech.

In a group study of wideband speech quality evaluation with 150 T-mobile subscribers, 70% of the participants preferred wideband speech. They reported they had a more relaxing atmosphere during a mobile phone call when they used wideband speech [17]. In another study published in the AT&T journal [15], listeners reported that it is more pleasant and less fatiguing to listen to 7 kHz bandwidth speech in comparison with telephone bandwidth speech. According to ITU-T Recommendation G. 107 [18], improvement of quality for wideband transmissions compared to narrowband transmissions is 29%.

1.2.2 Challenges

In addition to the abovementioned benefits of wideband speech, there are side effects associated with wideband telephony which need to be considered. Echo and noise are two challenging issues in wideband telephony. Users are more sensitive to echo and noise in wideband speech due to perceptual effects over a wider frequency range.

High frequency echo is more annoying for a few reasons and echo cancellers have a particularly difficult time in cancelling that; first, it falls into the area of audibility where the ear is more sensitive to sound [6]. Second, the loudness of echo in the extended frequency range of wideband speech will add to the loudness of echo in the narrowband frequency range and as a result the echo is perceived louder [6]. Third, high frequency echo is not masked as effectively as low frequency echo by one's own voice [6, 19].

Users are also more sensitive to wideband noise for the following reasons. The lower frequency limit of narrowband system, 300 Hz, filters out low frequency noise while wideband systems have the potential to transmit low frequency noise down to 50 Hz. The low frequency noise may contribute to upward spread of masking depending on its level. There is a similar, perhaps even worse scenario for the high frequency portion. According to [9], the high frequency portion of the speech spectrum is more prone to noise distortion than the low frequency part. The extra frequency range of wideband system allows more noise to be transmitted, which could potentially lead to more annoyance for the users. Second, loudness increases as the noise bandwidth increases. Since the total loudness is the sum of specific loudness at each critical band and bandwidth extension introduces more critical bands, the total loudness increases even when the overall sound level is held constant.

These challenges can be overcome by high performance signal enhancement algorithms, such as echo cancellation and noise suppression algorithms. However, these algorithms have been designed and optimized for the 300-3400 Hz bandwidth and cannot be directly used for wideband speech without any modifications [20]. Although preliminary investigations of noise reduction [20, 21] and echo cancellation [20, 22, 23] have been carried out for wideband application, a systematic investigation is currently lacking.

Another important issue of wideband telephony is the problem of interoperability between wideband and narrowband systems. The wide ranging use of wideband speech in telecommunication applications is expected soon and it is just a matter of time before a completely wideband transmission system is in place. During this transition period wideband and narrowband terminals will coexist and users will experience a noticeable quality difference between wideband and narrowband speech. Inconsistent loudness and quality would be two annoying parameters when one of the users is on a narrowband terminal and the other one is using a wideband terminal. Since the energy of narrowband signal spreads across fewer critical bands, it sounds quieter [6]. Users on narrowband terminal do not have any problem with this issue, because narrowband terminal filters wideband signal down to narrowband signal but for the users on the wideband terminal, this inconsistent loudness would be annoying. Experiencing wideband speech, users get more sensitive to narrowband speech impairments. Inconsistent quality would be annoying for the users on both wideband and narrowband terminals. It would be more noticeable when users are exposed to two levels of quality within a short period of time. For example if one user with wideband terminal is in a conference call where there are both narrowband and wideband terminals, switching between the talkers with different terminals makes the quality difference more significant. Users on narrowband terminal will also not be satisfied with the quality of narrowband speech because exposure to the wideband speech over time will have increased their expectation and make them more sensitive to the difference in quality.

During the transition time this quality gap between wideband and narrowband speech can be reduced by using bandwidth extension (BWE) techniques. Different methods have been proposed for BWE [11, 24, 25]. These methods try to take a narrowband signal and regenerate missing spectral content without any modifications to the existing transmission system. BWE cannot provide the same quality as wideband speech but it could be acceptable for the users on wideband terminals.

1.3 Speech Quality Evaluation

As mentioned before, the migration from narrowband to wideband telephony necessitates modifications to the signal processing blocks and it is imperative to understand the effect of these modifications on speech perception. Since this thesis concentrates on the aspects of speech quality, a more detailed description of speech quality assessment is given.

Speech quality can be measured in two ways: subjective and objective measures [26]. Subjective measures include the collection of ratings of speech quality by a group of listeners. While subjective measures have high face-validity, they are also expensive, and time-consuming. Therefore, objective (instrumental) measures which extract a metric of speech quality from the clean and processed signals are desired. Since objective test results are consistent and repeatable, the speech quality measurements conducted at different times and with different personnel and testing facilities can be directly compared. Before an objective metric of speech quality is used, its validity must be proven, i.e. there must be evidence that it exhibits significant correlation with the subjective speech quality scores, and can therefore be relied upon as a substitute for subjective ratings [26]. A brief overview of these two evaluation methods is given in the next two subsections.

1.3.1 Subjective Evaluation

Subjective measures are based on the perceptual ratings of processed speech by a listener or a group of listeners, who subjectively rank the quality of speech along a predetermined scale. Subjective measures are classified into utilitarian methods or analytical methods [26]. The utilitarian methods usually employ a unidimensional scale for reporting results. In contrast, the analytical methods always use a multidimensional scale for reporting the results and seek to identify the underlying psychological components that determine the perceived speech quality. The most widely used utilitarian subjective test is the absolute category rating (ACR) method which results in a mean opinion score (MOS), and this test was standardized by the International Telecommunications Union (ITU, the ITU-T Recommendation P.800) [27, 28].

In the ACR test, listeners rate the speech quality using a five-point scale, in which the quality is represented by five grades - excellent(5), good(4), fair(3), poor(2), and bad(1). Typically, the ratings are collected from a pool of listeners and the arithmetic mean of their ratings forms the MOS. The ITU-T P.800 [27] further describes the procedure for subjective

evaluation of speech degraded by different factors such as environmental noise, transmission error, talker echo etc. in telecommunication systems. In addition, specifications for talkers and listeners, speech material, and data collection process are detailed [27]. The ITU-T P.800 standard has been used to collect and create speech quality ratings databases, especially for the narrowband speech coding applications [29].

A different ITU-T recommendation, viz. P.835, specifically describes the subjective testing procedure for evaluating the quality of speech degraded by noise reduction algorithms [30]. The rating procedure in this specification was designed to reduce the listeners' uncertainty and confusion to score the quality of enhanced speech by using separate rating scales for evaluating "speech", "background noise" and "speech + background noise". Each trial in the rating task contains three tokens of the same processed speech sample and the listeners are instructed to listen to each of them and select one of the three (signal distortion, background noise, overall quality) five-point rating scale presented in the standard to register their opinion about the quality of that sample.

Four subjective testing methods for evaluating the performance of echo cancellers have been suggested in ITU-T recommendation P.831 [31]. The tests include: conversational tests, talking-and-listening tests, third party listening test type A and third party listening test type B.

In the conversational test, two parties converse over the system under test, during which they are asked to do some conversational tasks such as describing the position of a set of numbers on a picture to their partner. This test is the closest method for modeling the real-time interactions between the subjects; and also studying the effects of the impairments caused by these interactions. While the conversational test is the only test to model realistic conditions, it is both time-consuming and expensive to run. It is also hard to control the number and duration of the double-talk period [1](to simulate and study the double-talk impairments).

In the talking-and-listening test, a single subject must talk and listen simultaneously and then judge the quality of perceived speech, disturbances caused by echoes and quality of background noise transmission. Since there is no near-end subscriber during the test and subjects are not involved in any conversation they could focus more on the impairments caused by the echo of their own voice. In comparison with conversational test, this one is less realistic but easier to run.

[1]Double-talk period refers to when both users talk at once, i.e., there is speech on both of the two voice paths.

The third-party listening test differs from the conventional listening test, in that the listener can hear the signal from both end points while in the listening test, the listening point is at one end of the system under test. This test has two types: Type A: uses recordings made with Head and Torso Simulator (HATS) (according to Recommendation ITU-T P.58 [32]), one at each end of the connection. A subject, who plays the role of a third-party, listens to the recordings and rates the quality; the advantage of this test is that all the measurement conditions and test setup are controllable and repeatable, but this method is artificial in comparison with the first two methods; even though masking effect is being considered in this method but the naturalness of hearing one's own voice does not exist. Except for conversation related parameters such as delay which can be evaluated through the interactions between the two subjects, all other speech signal degradation can be covered and evaluated by this test. Type B is similar to the third-party listening Test A, but no HATS is used; this method has an easier recording procedure, but is more artificial.

It has been recommended that for evaluating echo cancellers performance, talking-and-listening tests, and listening only tests should not be done in isolation and should be followed by conversational test which involves interactions between subjects [31].

Recommendation ITU-R BS.1534-1 [33] describes another procedure for collecting subjective quality ratings. The test method, called "MUltiple Stimulus test with Hidden Reference and Anchors" (MUSHRA) was first proposed in [34] as an accurate and reliable method for audio quality evaluation of intermediate-quality signals.

In the MUSHRA protocol, all stimuli for one test condition are displayed on a user interface and the subjects are able to rate the quality of each sample in relation to the others as well as to the reference sample. The ratings will be done according to a five – interval Continuous Quality Scale (CQS) at which the scales are divided into five equal intervals with an internal numerical representation in the range of 0 to 100. The intervals are assigned the descriptors such as –Excellent –Good –Fair –Poor –Bad from top to bottom [34].

1.3.2 Objective Evaluation

Objective measures assess speech quality based on the extracted physical parameters from the speech signal or the system under test. Most objective measures comprise two blocks: feature extraction and feature mapping. The feature extraction block estimates the parameters that are representative of speech quality perception. These features are then mapped to a single quality score using the "cognitive" model which results in a higher degree of correlation with the subjective quality ratings.

Objective methods can be categorized based on three different criteria: 1) the measures which are used: parametric models (physical measures of the system) vs. signal-based methods (features of speech signal); 2) the information they need: from both sides of the system (end-to-end or intrusive method) or from one side (one-end or non-intrusive method); 3) the context they model (talking, listening or conversation) [35].

Intrusive models vs. non-intrusive models Intrusive (or end-to-end or with reference) models need both original (reference) and processed signals for assessing the quality of the speech. The reference signal is sent to the system or the algorithm under test and both reference and processed signals are employed by the model to predict the quality score. Non- intrusive (or single-ended or without reference) models evaluate the speech quality based only on the signal processed by the system or the algorithm under test. While using both reference and processed signals make the intrusive models more powerful for quality evaluation, there are many applications such as satellite communication and voice over IP networks where the reference signal is not readily available and intrusive methods are not applicable.

Signal-based models vs. parametric models Signal-based models and parametric models are different in the measures that they use for the evaluation. Signal-based models use the clean and processed speech signal (end-to-end model) or the processed signal only (single-ended model) to predict the quality of the system under test. Among the models with reference, the ones which are based on modeling the human auditory system are more attractive [36].

ITU-T P. 862 [37] has standardized a model called Perceptual Evaluation of Speech Quality (PESQ) as an end-to-end, signal based model for quality assessment of narrowband telephone networks and speech codecs. A wideband version of PESQ has been standardized in ITU-T P. 862-2 [38]. ITU-T recommendation P.563 [39] is the non-intrusive equivalent of PESQ that does not require a reference speech sample.

Parametric models use the parameters and physical measures of the system under test such as delay, SNR, echo, attenuation and so on, for assessing the voice quality. The state of the art parametric model is E-model, standardized by ITU in recommendation G. 107 [18] as a transmission planning tool. The model CCI (called clarity index), recommended in ITU-T P. 562 [40], is the equivalent of the E-model without reference. While parametric models can easily be placed in the network elements and terminals, they do not have the efficiency of the signal-based models in predicting the perceived speech quality [35].

The objective measures can also be categorized based on the context they can model: lis-

tening, talking and conversational test [35]. For example both PESQ and P. 563 have been standardized for estimating subjective quality obtained in listening-only tests. While the E-model can be used in both listening and conversational context, there is no standardized signal-based model for the speech quality evaluation in conversational and talking-quality context.

Appel and Beerends [41] developed an objective perceptual talking-quality measure, called the perceptual echo and sidetone quality measure (PESQM). Guéguin et al. [35] used this model along with PESQ and delay parameter to develop a signal-based measure for the quality evaluation in conversational context.

1.3.3 Subjective Measure vs. Objective Measures

An ideal objective metric should be able to predict the subjective quality scores with high accuracy. Various statistics can be used to evaluate the performance of the objective metrics. The two most common ones are Pearson's correlation coefficient and the standard deviation of error [36].

The degree of linear relationship between the predicted scores by the objective model (y) and the actual quality scores (q) can be obtained using Pearson's correlation coefficient:

$$\rho = \frac{\sum_{i=1}^{N} (q_i - \tilde{q})(y_i - \tilde{y})}{\sqrt{\sum_{i=1}^{N} (q_i - \tilde{q})^2 (y_i - \tilde{y})^2}} \tag{1.1}$$

The standard deviation of the error shows the average of the variability of the subjective scores about the regression line:

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{N} (y_i - q_i)^2}{N}} \tag{1.2}$$

which also can be written as

$$\sigma_e = \sigma_q \sqrt{1 - \rho^2} \tag{1.3}$$

where σ_q is the standard deviation of the subjective scores [36].

A good objective metric should yield a high correlation value and a small value of σ_e.

1.4 Problem Statement and Thesis Scope

There are two main objectives of the research presented in this thesis.

The first is to evaluate the performance of narrowband noise reduction algorithms when deployed in wideband application and improve their performances with fine-tuning their parameters. For this purpose, and because of the lack of wideband databases containing clean and enhanced speech processed by different noise reduction algorithms, a custom database was first created. Using the subjective quality scores of speech stimuli in the custom database, an objective model was developed and validated. Since the quality scores predicted by the objective model were highly correlated with the subjective scores, it was used as a substitute for subjective quality scores for fine-tuning wideband noise reduction algorithms and improving their performance for wideband application. A paired comparison subjective test was performed to compare the performance of the algorithms before and after the updates.

The second goal is to evaluate the effect of echo and delay on the wideband speech quality in both listening and conversational context. For the listening context, a database containing speech samples corrupted by different amount of echo, delay and echo path models was created and subjective quality ratings were collected by presenting the samples to the subjects. The performance of the objective model developed for performance assessment of the noise reduction algorithm, was evaluated in terms of predicting the quality of speech samples corrupted by echo. The model was also used for evaluating the performance of an echo canceller and echo suppressor.

One of the main limitations of listening only tests is that it cannot be used for studying the effects of conversational related parameters such as delay and double-talk. As such, signal processing algorithms such as the echo canceller algorithm should also be evaluated during a real conversation. To facilitate the conversation test, a realtime simulator for telephone conversation was developed. This simulator was later used to investigate the effect of delay and echo in the conversational context, by collecting subjective speech quality scores and by predicting these scores through the objective model.

1.5 Thesis Organization

The thesis is prepared in the monograph format including six chapters:

In Chapter 2, the structure of a typical intrusive signal based objective speech quality measure is described. For the feature extraction block, an overview of some of the objective metrics of speech quality published in the literature with more focus on auditory-based models is presented. Two multivariate regression models as candidates for the feature mapping block are described.

In Chapter 3 the performance of several noise reduction algorithms intended for wideband telephony was compared and evaluated subjectively. For this purpose, a customized wideband noise reduction database containing speech samples corrupted by three types of background noises at three SNR levels, along with their enhanced versions was created. Using the collected subjective scores, the performance of several well-known auditory-based objective metrics was evaluated in predicting the wideband speech quality. Furthermore, the performance of a cognitive model called Bayesian Multivariate Regression Model (BMLS) was evaluated in mapping the features extracted by the auditory-based models into the true quality scores. The combined model which resulted in the best performance was used for optimizing the parameters of a few noise reduction algorithms. The performance of the fine-tuned algorithms before and after the updates was evaluated and compared using paired comparison subjective test.

In Chapter 4, the effect of impairments caused by echo and delay on the quality of wideband speech was investigated in the listening context. A database of wideband speech samples corrupted by echo, delay and different echo path models was created for this purpose. The effect of echo, delay and the shape of the frequency response of a few echo path models was evaluated subjectively and objectively. The performance of an echo canceller and echo suppressor were also evaluated using the objective metric.

In Chapter 5, the effect of delay and echo in the wideband conversational context was evaluated. A software module to simulate a real-time conversation was developed for this purpose. Speech materials were also designed for the test. Pairs of subjects conversed over the simulated telecommunication system and rated the quality of their conversation. the quality scores were used for subjective evaluation of the effect of echo and delay. The signal-based objective metric developed in chapter 3, along with the delay parameter were used as the objective measure for predicting the quality of the conversation.

Chapter 6 of this thesis is an overall conclusion with suggestions for the future work.

Chapter 2

Speech Quality Evaluation

2.1 Introduction

This chapter provides the methodological details in objective assessment of speech quality. Objective speech quality metrics are typically derived through two processing blocks: feature extraction and feature mapping. The feature extraction block estimates the parameters that are representative of speech quality perception, which are then mapped to a single quality score using the feature mapping or "cognitive" model, resulting in a higher degree of correlation with the subjective quality ratings. The block diagram of a typical objective speech quality model is shown in Figure 2.1.

Good performance from both feature extraction and feature mapping blocks is necessary to predict subjective speech quality scores with acceptable accuracy. For example, it has been shown that features extracted by auditory models rather than signal-based features are more correlated with the subjective quality scores [42, 43].

The improvement can also be obtained from the combination of several objective metrics using a cognitive model [44, 45]. In other words, a feature vector that is inclusive of different parameters derived from temporal, spectral, and perceptual modeling may result in better performance after feature mapping, as each parameter captures different aspects of degradations in speech quality.

In the next section, an overview of some of the objective metrics of speech quality published in the literature is provided. This is followed by Section 2.3, where the focus is on multivariate regression functions as potential feature mappers for speech quality estimation.

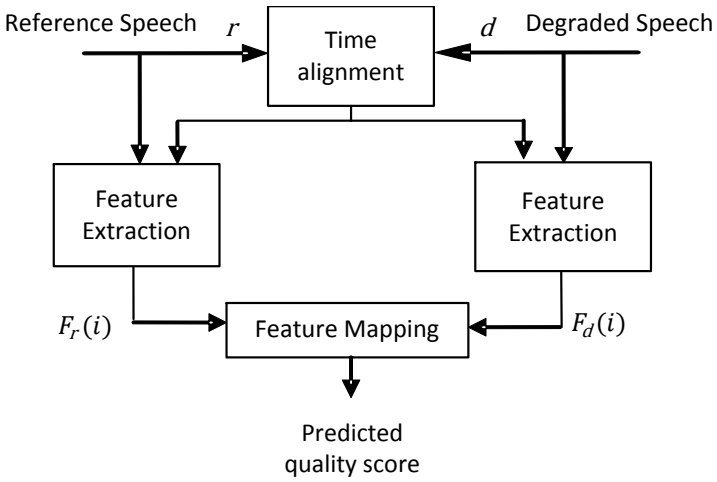

Figure 2.1: Block diagram of an objective speech quality measure

2.2 Features for Speech Quality Assessment

Features proposed in the literature for estimating speech quality can be categorized under four main groups [26, 36, 46]: 1) time domain based measures, such as the Signal-to-Noise Ratio (SNR) and Segmental Signal-to-Noise Ratio (SNRseg), 2) Linear Predictive Coefficients (LPC) based measures, such as Log-Likelihood Ratio (LLR) and Itakura Saito (IS) measure, 3) spectral domain based measures such as the Spectral Distance (SD) and Log Spectral Distance (Log SD), and 4) perceptual or auditory-based measures, where the signal processing involved in the peripheral auditory system is explicitly included in the feature computation. Weighted Spectral Slope (WSS) and PESQ are two examples of the last category.

There is substantial evidence that the auditory model based features outperform time-domain or LPC-based features. As such, they are discussed in more detail later in this chapter. In addition, a brief description of two LPC-based objective measures is given for comparative purposes.

2.2.1 LPC-based Objective Measures

Both LLR and IS are based on the differences between the all pole models of the reference and distortion speech waveforms. Theses two model assume that over each short-length frame, speech can be represented by a p^{th} order all pole model.

Log-likelihood Ratio (LLR)

The LLR measure is defined as:

$$d_{LLR}(a_r, \bar{a}_d) = \log \frac{\bar{a}_d^T \mathbf{R}_r \bar{a}_d}{a_r^T \mathbf{R_r} a_r} \tag{2.1}$$

where, $a_r^T = [1, -\alpha_r(1), -\alpha_r(2), \cdots, -\alpha_r(p)]$ are the LPC coefficients of the reference signal, $\bar{a}_d^T = [1, -\alpha_d(1), -\alpha_d(2), \cdots, -\alpha_d(p)]$ are the coefficients of the distorted signal and R_r is the $(p+1) \times (p+1)$ autocorrelation matrix of the reference signal.

Itakura Saito (IS)

The IS measure is defined as follows:

$$d_{IS}(a_r, \bar{a}_d) = \frac{G_r}{\bar{G}_d} \frac{\bar{a}_d^T \mathbf{R_r} \bar{a}_d}{a_r^T \mathbf{R_r} a_r} + \log\left(\frac{\bar{G}_d}{G_r}\right) - 1 \tag{2.2}$$

where G_r and $\bar{G}_d$ are the all-pole gains of the reference and distorted signals, respectively.

2.2.2 Speech Quality Metrics Based on Auditory Models

Weighted Spectral Slope (WSS)

The WSS distance measure computes the weighted difference between the spectral slopes of reference and processed speech spectra in each frequency band. The band-specific spectral slope differences are weighted according to, first, whether the band is near a spectral peak or valley and, second, according to whether the peak is the largest peak in the spectrum. The WSS measure is finally computed for each frame of speech as:

$$WSS = \sum_{k=1}^{L} W(k)(S_r(k) - \bar{S}_d(k))^2 \tag{2.3}$$

where, $W(k)$ is the weight for band k, $S_r(k)$ and $\bar{S}_d(k)$denote the spectral slopes of the reference and distorted signals of the k^{th} band and L is the number of critical bands which is used.

It must be noted here that, WSS models only one stage of auditory modeling, viz. the critical band auditory filterbank. This is in contrast with more comprehensive auditory

models that incorporate several stages of the auditory processing, and these are discussed in detail below.

Perceptual Evaluation of Speech Quality (PESQ)

PESQ is a popular speech quality estimation method standardized by the ITU for both narrowband [37] and wideband [38] telephony applications. In PESQ, a set of features are extracted from the test signal and its corresponding clean reference signal, and are subsequently compared in a perceptual space. PESQ computational procedure involves three modules: the time alignment module, the perceptual model of auditory periphery, and the cognitive model. The time alignment module undertakes a multi-step normalization of delay between the test and reference signals. The perceptual model incorporates a time–frequency analysis procedure, wherein the test and reference signals are divided into 32 ms frames with successive frames overlapped by 50% and transformed to the frequency domain using the short-time Fourier transform. Two computational steps that reflect human auditory perception are then applied to the time–frequency cells: (a) transformation of the linear frequency axis to the Bark scale, which accounts for the finer frequency resolution at lower frequencies than higher frequencies, and (b) transformation of the amplitude values to "loudness" values according to Zwicker's loudness formula [37]. The resulting loudness densities from the test and reference signals are then compared and differences in the densities ("disturbance density") are aggregated together to give rise to the PESQ score.

Loudness Pattern Distortion (LPD)

As described in the previous section, the ITU standardized PESQ is based on the Zwicker's loudness model. Recently, Moore and Glasberg (M–G) [47] developed an enhanced auditory model that better matches the data from psychoacoustical experiments.

In addition, the M–G model can be used to better explain how equal-loudness contours change as a function of level, why loudness remains constant as the bandwidth of a fixed-intensity sound increases up to the critical bandwidth, and the loudness of partially masked sounds. Speech quality metrics derived using the M–G model have been shown to correlate well with behavioural data evaluating speech coding algorithms [46].

The block diagram of feature extraction based on the M–G model is shown in Figure 2.2. Both the reference speech r and distorted speech d are separately analyzed by identical operations, leading to the loudness patterns, N'_r and N'_d, respectively. The process is started

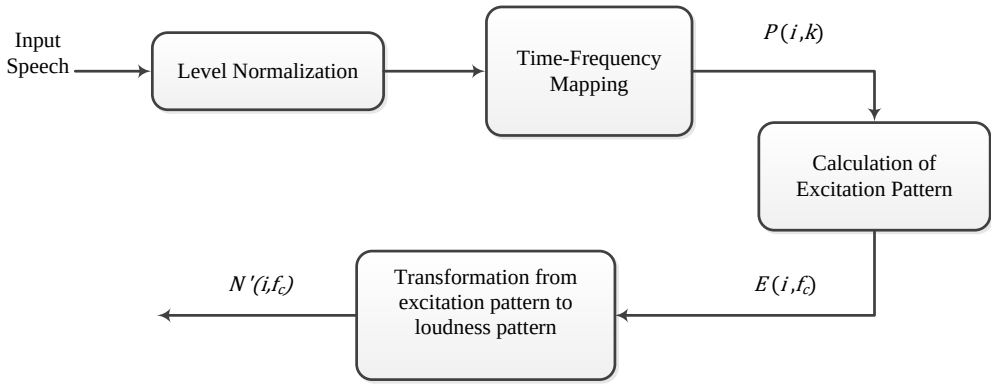

Figure 2.2: Block diagram of loudness pattern calculation according to the Moore-Glasberg model.

by normalizing the speech signal, segmenting it into frames and mapping the frames to the frequency domain. The individual frame spectra are passed through a bank of bandpass filters simulating the auditory filterbank. The shape of the auditory filters at different centre frequencies and levels was derived in Glasberg and Moore [48]. The output level of these filters as a function of centre frequencies is called the excitation pattern. Using the speech power spectral density and the auditory filters, the excitation pattern is computed as:

$$E(i,f_c) = \int_0^{\infty} \varphi(f,f_c,P)P(i,f)df \tag{2.4}$$

where $P(i,f)$ is the power spectral density of ith frame, $\varphi(f,f_c,P)$ is the ro-ex auditory filter with the centre frequency f_c. From the excitation pattern, the loudness pattern can be calculated by considering three different cases as follows,

$$N'_{SIG}(i,f) = \begin{cases} C\left(\dfrac{E_{SIG}(i,f)}{1.04\times 10^6}\right)^{1.5} & \text{if } E_{SIG}(i,f) > 10^{10} \\ C\left[E_{SIG}(i,f)G(i,f)+A(i,f)\right]^{\alpha(i,f)} - A(i,f)^{\alpha(i,f)} & \text{if } E_{THRQ}(f) \le E_{SIG}(i,f) \\ & \quad \& \; E_{SIG}(i,f) \le 10^{10} \\ C\left(\dfrac{2E_{SIG}(i,f)}{E_{SIG}(i,f)+E_{THRQ}(f)}\right)^{1.5} \times & \text{if } E_{SIG}(i,f) < E_{THRQ}(f) \\ \left(\left[E_{SIG}(i,f)G(i,f)+A(i,f)\right]^{\alpha(i,f)} - A(i,f)^{\alpha(i,f)}\right) & \end{cases} \tag{2.5}$$

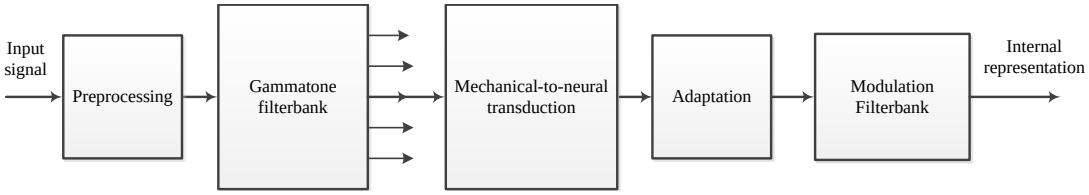

Figure 2.3: Block diagram of PEMO-Q computation.

where f and i are frequency and frame indices, respectively. E_{THRQ} is peak excitation evoked by a sinusoidal signal at absolute threshold,

$$T_{THRQ}(f) = 1.4 + 0.4 \times 10^{0.3\left(\frac{f}{kHz}\right)^{-0.8}} \tag{2.6}$$

C is a constant with value of 0.047, A and α are constant for frequencies of 500 Hz and above, and G is the low level gain of the cochlear amplifier at a specific frequency relative to the gain at 500 Hz and above (which is assumed to be constant),

$$A(f) = 2.8 + \frac{2}{0.1 + G(f)^{0.25}} \tag{2.7}$$

$$\alpha(f) = 0.171 + \frac{0.032085}{0.1 + G(f)^{0.25}} \tag{2.8}$$

$$G(f) = \frac{E_{THRQ}(500Hz)}{E_{THRQ}(f)} \tag{2.9}$$

Finally, the loudness pattern distortion (which is used as the speech quality measure) is obtained by:

$$X(f_c) = \sqrt{\frac{\sum_{i=1}^{I}\left[N'_r(i,f_c) - N'_d(i,f_c)\right]^2}{\sum_{i=1}^{I}\left[N'_r(i,f_c)\right]^2}} \tag{2.10}$$

where N'_r and N'_y are the loudness of reference and degraded speech respectively, and I denotes the number of frames.

Perceptual Model – Quality Assessment (PEMO-Q)

The M–G model described above was proposed for estimating loudness perception of stationary sounds. Alternatives to the M–G model which take into account the temporal and spectral masking phenomena in human auditory system have been proposed in the literature [49, 50, 51, 52, 53]. One of these models, named PEMO-Q [53], has been already used for predicting the quality of wideband speech samples degraded by small impairments caused by audio codecs; the predicted scores exhibited a high degree of correlation with the subjective quality ratings [53].

The block diagram of PEMO-Q structure is shown in Figure 2.3. Before being analyzed by the auditory model, both the reference and degraded signals are pre-processed by removing their time delay and level differences. Then using a Voice Activity Detector (VAD), the silent intervals of the reference signal as well as the corresponding sections of the degraded signals are removed. The auditory model in PEMO-Q is based on the model developed by Dau et al. [51, 52] and is initiated by dividing the input speech into critical frequency bands by a linear fourth-order gammatone filterbank. This filterbank can model level- and frequency-dependent compression as well as bandpass characteristic of the basilar membrane (BM). The centre frequencies of the filterbank were separated by one equivalent rectangular bandwidth (ERB) and each filter also has a bandwidth of one ERB. The output of the gammatone filterbank has multiple channels, and simulates the temporal output activity at different frequencies. Each channel is processed separately, starting with half-wave rectification followed by a first-order lowpass filter with a cutoff frequency of 1 kHz. This step models the transfer of the mechanical vibrations of the BM into inner hair cell receptor potentials. In the next stage, a chain of five simple nonlinear circuits, with different time constants are applied on the signal. Each circuit consists of a lowpass filter and a division operation, modeling the adaptation nature of the peripheral auditory system in adjusting its gain with changes in the level of input signal. In the final stage, the envelope signal is analyzed by a bank of modulation bandpass filters. The output of this stage is referred to as the "internal representation" and has three dimensions: time, frequency and modulation frequency.

The cross correlation coefficient of the internal representations of the reference and the test signal is used for calculating the quality index. For this purpose, first the internal representations of the reference (r_{tmf}) and degraded (d_{tmf}) signals are calculated and then for each modulation channel, the linear cross correlation coefficient of two $K \times L$ matrices is given by:

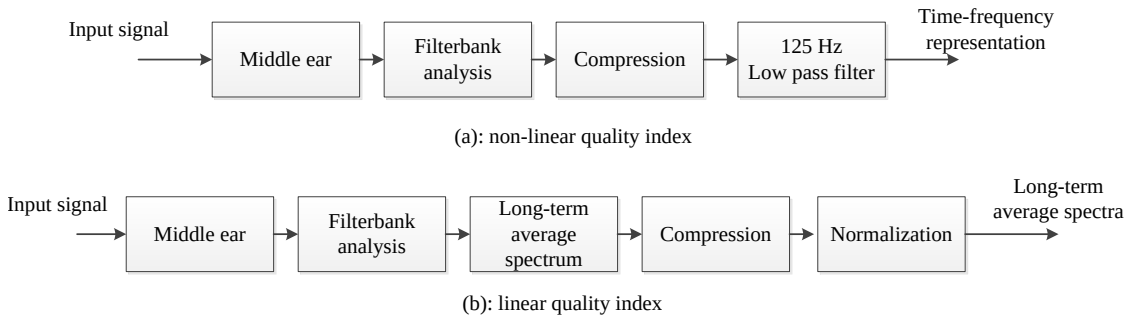

Figure 2.4: Block diagram of HASQI. (a) nonlinear quality index (b) linear quality index.

$$\rho_{Pm} = \frac{\sum_{t,f=1}^{K,L}(r_{tf}-\bar{x})(d_{tf}-\bar{d})}{\sqrt{\sum_{t,f}\left(r_{tf}-\bar{r}\right)^2\sum_{t,f}\left(d_{tf}-\bar{d}\right)^2}} \tag{2.11}$$

where K, L represent the number of time samples and frequencies, and $\bar{r}$ and $\bar{d}$ denote the mean values respectively. Finally, the perceptual quality measure (PSM) is computed as $PSM = \sum_m w_m \rho_{Pm}$, with $w_m = \frac{\sum_{t,f=1}^{K,L} d_{tf}^2}{\sum_{t,f,m'=1}^{K,L,M} d_{tfm'}^2}$ where M is the number of modulation channels.

The Hearing-Aid Speech Quality Index (HASQI)

The HASQI model [54] has been developed recently for predicting the speech quality ratings by both normal hearing and hearing impaired listeners. The model takes into account the effect of noise, nonlinear distortion as well as linear processing on the speech quality. HASQI is a combination of two quality indexes; while the noise and nonlinear index captures the differences in the short-term signal envelope behaviour, the linear index focuses on the linear filtering and its effects on the long-term average speech spectrum. The product of these two indexes results in the combined quality index [54].

The block diagram of HASQI structure is shown in Figure 2.4; part (a) and (b) show the procedure of calculating nonlinear and linear quality indexes respectively. In part (a), the input signal is first filtered by the middle ear filter and then analysed by the gammatone auditory filterbank, in a manner similar to the PEMO-Q model. The outputs of the filterbank are processed by a compression block, which simulates the level-dependent compressive behaviour of the basilar membrane. The compressed filterbank outputs are then processed through a 125 Hz lowpass filter to extract the envelopes. The smoothed envelopes thus computed from the reference and processed speech samples are subsequently compared in the cepstral domain. The envelopes are fitted with a set of six cepstral bases functions,

and the degree of correlation for each fitted basis function between the reference and processed envelopes is calculated and is subsequently mapped to the nonlinear quality index. A second-order polynomial was used as the mapping function.

Figure 2.4-part(b) shows the procedure used for calculating the linear quality index, which is based on the sound quality metric developed by Moore and Tan [55] except that a different cochlear model is used in HASQI. Middle ear filter and the filterbank analysis are same as part (a). The outputs of the filters are averaged across time to derive the long-term average spectrum. After the compression stage, the signal is converted back to the linear amplitude and normalized to have an *RMS* value of 1 when summed across the auditory bands; this normalization removes the effect of the signal amplitude on the speech quality. Both original and degraded signals are processed by the structure given in part (b); having the normalized original and degraded signal spectra, the difference in the spectra $d_1(k)$, as well as the difference in the spectral slopes $d_2(k)$ is calculated. These two values are then mapped to the subjective quality ratings using an MMSE (minimum mean square error) linear regression and result in the linear quality index.

2.3 Mapping Function

The cognitive model involves determining the relationship between n observations on the speech quality estimations, $y = (y_1, \ldots, y_n)'$, and their corresponding features, $X = (x_1, \ldots, x_n)'$, i.e.: $y_i = f(x_i) + \varepsilon_i$, where ε_i has normal probability distribution, $N(0, \sigma^2)$. The true regression function "f" is unknown and needs to be approximated.

There are a variety of methods for multivariate regression analysis. Generalized additive models, neural networks, MARS (Multivariate Adaptive Regression Spline) and BMLS (Bayesian Multivariate Linear Splines) are some of them [56].

Each method has some advantages and disadvantages. For example while the prediction ability of neural network methods is high, they are hard to interpret [57].

MARS and BMLS are both interpretable and flexible. While MARS is suitable for data with moderate dimensions $3 < N < 20$ [58], BMLS is an alternative for MARS for high-dimension regression problems [56, 59].

2.3.1 Multivariate Adaptive Regression Spline (MARS)

MARS, introduced by Friedman [58], is one of the most popular nonlinear regression methods as it is highly flexible and easily interpretable. The model is built as a product of spline basis functions where the number of basis functions and other parameters such as knot location and product degree are automatically determined by the data.

$$f(x) = \beta_0 + \sum_{i=1}^{k} \beta_i B_i(x) \tag{2.12}$$

where $B_i(x)$ is:

$$B_i(x) = \prod_{i=1}^{J_i} [s_{ij}(x_{w_{ij}} - t_{ij})]_+, \;\; i = 1, 2, \cdots, k \tag{2.13}$$

where $[.]_+ = max[0, .]$, J_i is the degree of interaction of basis function B_i, the s_{ij} are the sign indicators taking values $\pm$ 1, the t_{ij} are knot points and the w_{ij} give the index of the predictor variables which is being split on the t_{ij}. Also w_{ij} are constrained to be distinct, so each predictor only appears once in each interaction term. Hence, a MARS basis is just a product of univariate linear spline terms; more details about the model can be found in [58]. It is worth pointing out here that the maximum number of basis functions should be given to the model by users.

2.3.2 Bayesian Multivariate Linear Splines (BMLS)

The basic principle of Bayesian approach is to calculate the conditional probability distribution of the unobserved variables of interest, given the observed data. It means that the posterior predictive distribution of new output y_{n+1} must be calculated for the new input x_{n+1} given the training data set D, i.e.,

$$p\left(y_{n+1}|x_{n+1}, D\right) = \int p\left(y_{n+1}|x_{n+1}, W\right) p\left(W|D\right) \, dW \tag{2.14}$$

where W denotes all the model parameters and hyper-parameters [1]of the prior structures, and $p\left(W|D\right)$ represents the posterior probability of the parameters of the model f given the

[1]Hyper-parameters are the parameters of a prior distribution. This name is used to distinguish between the parameters of the model and prior distribution.

training data set D. The estimation of speech quality can be obtained by:

$$\hat{y}_{n+1} = E\left(y_{n+1}|x_{n+1}, D\right) = \int f\left(x_{n+1}, W\right) p\left(W|D\right) dW \tag{2.15}$$

In real regression analysis, there is no single model which can predict the true relationship between the inputs and the outputs. While classical methods try to select the best model with optimising the parameters of the model, the Bayesian approach integrates out the uncertainty between the parameter values and averages over models with different number of basis functions, where each model is weighted by the posterior probability of its parameters. For taking into account the uncertainty between models M of different dimension (e.g., the number of basis functions), the expectation in Equation 2.15 is written as,

$$\hat{y}_{n+1} = \sum_{k=0}^{K} \int f\left(x_{n+1}, W_k, M_k\right) p\left(W_k|D, M_k\right) p\left(M_k|D\right) dW_k \tag{2.16}$$

where $M = \{M_0, \ldots, M_k\}$ is the set of entertained models and $p\left(M_k|D\right)$ is the posterior distribution of model M_k, obtained by using prior distribution of model and Bayes rule:

$$p\left(M_k|D\right) = \frac{p\left(D|M_k\right) p\left(M_k\right)}{p\left(D\right)} \tag{2.17}$$

Let M_k denote a typical model, then using the Bayes rule prior distributions on the model $p(M_k)$ are updated to posterior distributions $p(M_k|D)$.

Since the posterior distribution has a complex form, the integral in equation (2.14) cannot be calculated using analytical methods. Instead, a reversible jump MCMC (Markov Chain Monte Carlo) sampling strategy [60] was used to approximate the integral by drawing samples from the joint probability distribution of all the model parameters, $p(W|D)$, and then approximating the integral in (2.14) by:

$$I \approx \frac{1}{N - n_0} \sum_{t=n_0}^{N} f\left(W_t\right) \tag{2.18}$$

where N is the total number of the generated samples, $W_1, \ldots, W_N$ are draws from the posterior distribution of W and n_0 is a "burn-in" period. To give the algorithm a chance to converge to $p(W|D)$, the samples (i.e. W_t) from the first few iterations of the algorithm, known as the burn-in period, are discarded and after that convergence is assumed and every k^{th} (here k=5) sample is saved as the valid sample for calculating Equation 2.18; in this

way the correlation between the successive samples is removed and the generated samples of the models are less dependent [59, 61].

In BMLS model, piecewise linear planes are used as basis functions and the regression function can be written as,

$$\hat{f}(x_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \left(\mathbf{x_i}.\mu_\mathbf{j}\right)_+ \tag{2.19}$$

where $\beta = (\beta_0, \ldots, \beta_k)^{'}$ and $\mu_\mathbf{j} = (\mu_{j0}, \ldots, \mu_{jp})$ are regression coefficient and basis parameter respectively; p is dimension of the predictors, $a.b$ is inner product and $(a)_+ = max(0,a)$; $\mathbf{x_i}.\mu_\mathbf{j}$ is a truncated linear plane which its position and orientation is determined by parameter $\mu_\mathbf{j}$. In matrix format,$\hat{f}(x_i)$ can be written as: $\hat{f}(x_i) = B\beta + \varepsilon$ which B is the $n \times (k+1)$ matrix:

$$B = \begin{pmatrix} 1 & (x_1.\mu_1)_+ & \cdots & (x_1.\mu_k)_+ \\ 1 & (x_2.\mu_1)_+ & \cdots & (x_2.\mu_k)_+ \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n.\mu_1)_+ & \cdots & (x_n.\mu_k)_+ \end{pmatrix}.$$

To introduce covariate selection to the basis functions, some of the elements of μ_{j-0} (μ_j except the first element) are set to zero, and make the plane perpendicular to the corresponding covariates. To determine the number and place of the non-zero elements in μ_{j-0}, two new parameters are introduced: $\gamma_j = \left(\gamma_{j1}, ..., \gamma_{jp}\right)$ and z; $\gamma_{jd} = 1$ if the d^{th} element of μ_{j-0} is non-zero and vice versa, and $z = \sum_{d=1}^{p} \gamma_{jd}$ which shows the number of non-zero elements in μ_{j-0} [56].

Function "f" can be uniquely determined by the number of basis function k, the position vector μ, γ, the number z, the output coefficients β and the regression variance σ^2, which is the variance of the noise. Therefore BMLS is parameterised by $W = (M_k, w)$ where M_k is defined to include the number and location of the basis functions, M_k=(k, μ_k, γ, z), and w includes (β, σ^2).

In a Bayesian network, the posterior probability densities of these parameters are of interest given the data set. The posterior density is given as a combination of likelihood and prior. The Bayesian approach is based on three basic steps [59]:

1) Assigning prior distributions to all the unknown parameters, $p(W)$.

2) Calculating the likelihood of the training data, using the given parameters, $p(D|W)$.

3) Determining the posterior distributions of the parameters, using Bayes rule.

Assigning prior distributions to all the unknown parameters

The prior density is used to represent information about the unknown parameters and incorporate inferences for simpler models or smoother model outputs. The unknown parameters in the model are number of basis functions k, knot position μ, γ, z, the set of coefficients $\beta = (\beta_1, ..., \beta_k)'$ and the regression variance σ^2.

It is preferred to choose mathematically convenient forms of prior distribution which result in computationally tractable posterior distribution. This goal is achieved through the use of conjugate prior distribution. For BMLS model, the conjugate choice of prior for β and σ^2 is the normal inverse-gamma (NIG).

$$\begin{aligned} p\left(\beta, \sigma^2 | M_k\right) &= p\left(\beta | \sigma^2, M_k\right) p\left(\sigma^2 | M_k\right) \\ &= N(\beta | 0, \lambda^{-1}\sigma^2 I) InvGamma(\sigma^2 | a, b) \\ &= \frac{b^a}{(2\pi)^{\frac{k}{2}} |\lambda^{-1}|^{1/2} \Gamma(a)} \left(\sigma^2\right)^{-\left(a+\left(\frac{k}{2}\right)+1\right)} \\ &\quad \times exp\left[-\left(\beta'\lambda\beta + 2b\right)/2\sigma^2\right] \end{aligned} \tag{2.20}$$

where, a and b are parameters of inverse gamma distribution and λ is the precision (inverse variance) of the normal distribution.

Uniform prior is assigned on $z, U[1, 2, ..., Z]$, which Z is the maximum allowed interaction. A uniform prior distribution is also assigned on γ conditioned on z as well as μ conditioned on γ and z. The final prior distribution is adopted for k, the number of basis function; as there is no information available about this number, uniform distribution from zero to the number of training data is assigned to that, $p(k) = U(0, ..., n)$.

To summarize, the joint prior distribution on the model parameters is:

$$p\left(k, \beta, \mu, \sigma^2, z, \gamma\right) = p\left(\beta | \sigma^2, k\right) p\left(\sigma^2\right) p\left(\mu | z, \gamma, k\right) p\left(\gamma | z, k\right) p\left(z | k\right) p\left(k\right) \tag{2.21}$$

which $p\left(\beta | \sigma^2, k\right) p\left(\sigma^2\right)$ is set to the normal-inverse gamma and the rest to uniform distribution.

Calculating the likelihood of the data, given the parameters

Assuming that the noise term takes a normal distribution, the likelihood $p\left(D|\beta,\sigma^2,M_k\right)$ or alternatively $p\left(Y|X,\beta,\sigma^2,M_k\right)$ can be written as,

$$\begin{aligned} p\left(D|\beta,\sigma^2,M_k\right) &= N\left(f_{M_k}(x),\sigma^2 I\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} exp\left[-\frac{(Y-B\beta)'(Y-B\beta)}{2\sigma^2}\right] \end{aligned} \tag{2.22}$$

and in log format, the following log-likelihood of the observed data is obtained:

$$L\left(D|\beta,\sigma^2,M_k\right) = -nlog\sigma - \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[y_i - f\left(x_i,\beta,\sigma^2,M_k\right)\right]^2 + constant \tag{2.23}$$

Determining the posterior distributions of the parameters

The posterior distribution of β and σ^2 has standard format, thanks to the conjugate prior distributions. The posterior distribution of the other parameter is complex and dimensional varying (the number of basis functions, k, is one of the parameters which is unknown) and cannot be calculated analytically.

As it was mentioned earlier, a reversible jump MCMC method is used for sampling from the posterior distribution. This method is a generalization to the Metropolis-Hastings algorithm [59] with introducing a number of other possible move types surrounding a change in dimension of the density.

The sampling algorithm starts with one basis function with unity values for all the input features; at each iteration, the sampler can suggest one of the three following proposals:

Birth - Adding a basis function to the model

Death - Removing one of the basis functions(Death move is not proposed when $k = 0$)

Move - Changing the parameter set of one of the existing basis functions

Assuming $M = (\alpha,\beta,k,\mu,\gamma,z)$ shows the current state of the model and the sampler propose a model with parameters $M^\star = (\alpha^\star,\beta^\star,k^\star,\mu^\star,\gamma^\star,z^\star)$, the proposal will be accepted by probability:

$$S(M,M^\star) = min\left(1,\frac{p\left(D|k^\star,\mu^\star,\gamma^\star,z^\star\right)}{p\left(D|k,\mu,\gamma,z\right)}\right) \tag{2.24}$$

Table 2.1: A summary of the described auditory-based models.

Objective measures	**Features**	**Mapping function**	**Comments**
WPESQ	Loudness pattern	$y = 0.999 + \frac{4.999-0.999}{1+\exp(-1.3669x+3.8224)}$	Loudness pattern is calculated on the Bark scale domain, using Zwicker's law.
LPD	Loudness pattern	Distortion function given in Formula 2.10	Loudness pattern is calculated on ERB scale, based on Moore-Glasberg auditory model.
PEMO-Q	Internal representation (time, frequency, modulation frequency)	The weighted sum of the correlation coefficients of the internal representation of the clean and enhanced signals per modulation channel	The envelope signal analysed by a bank of modulation filterbank resulted in internal representation of the signal.
HASQI	Signal envelope and long-term average speech spectrum	A second-order polynomial was used as the mapping function for calculating the non-linear index. An MMSE linear regression function was the mapping function for calculating the linear index. The product of the linear and nonlinear indexes resulted in HASQI score.	The envelopes are fitted with a set of six cepstral bases functions, and the degree of correlation for each fitted basis function between the reference and processed envelopes is used as the features of nonlinear index. The difference in the spectra, as well as the difference in the spectral slopes of the clean and enhanced signals was the features of the linear index.

which $p(D|k^\star,\mu^\star,\gamma^\star,z^\star)$ and $p(D|k,\mu,\gamma,z)$ are likelihood of the data given the models. The fraction is known as Bayes factor and is a well-known model selection criterion for controlling the model complexity [56].

The sampling process is iterated until enough samples have been drawn from the target distribution, specified by parameter "sample". An initial portion of the chain (i.e. burn-in) is discarded to ensure convergence, and after that every fifth of the next samples were stored to be used by Equation (2.18) for the prediction.

2.4 Summary

This chapter presented a review of a few objective quality measures and especially the ones which involve the modelling of peripheral auditory system. A summary of these models is given in Table 2.1. Further, two cognitive models were introduced as candidates for being used as the "feature mapping block" of the objective measure. These models can be used to combine the objective metrics to build composite objective measures that encompass the advantages of all its individual components. The performance of the objective models pre-

sented in this chapter for evaluating the quality of wideband speech has not been explored so far. This very evaluation and validation forms the focus of next three chapters.

Chapter 3

Wideband Noise Reduction and Speech Quality

3.1 Introduction

The use and prevalence of hands free and mobile communication devices have increased over the past few years. In such communication situations, speech is more susceptible to background noise and consequently the quality of speech will be degraded. Therefore, having a noise reduction system as a preprocessor is vital to render the voice communication more pleasant and natural. Several noise reduction algorithms have been and continue to be proposed for telecommunication applications [36]. Benchmarking these noise reduction algorithms is imperative not only to rank order the algorithm performance, but also to fine-tune the algorithm parameters such that the performance is optimized.

Substantial body of research exists on speech quality prediction [26, 62], but only a few studies investigated the application of objective quality metrics to the assessment of noise reduction performance [42, 44, 45, 63]. In order to undertake this evaluation, a database containing subjective speech quality ratings of noise reduction performance is essential. Hu & Loizou [64] created one such database when they conducted a systematic investigation of the narrowband noise reduction algorithms following the P.835 [30] procedure. They reported subjective scores for a noise reduction database, which contained male and female speech samples corrupted by four different noise types (car, babble, street, and train), corrupted by two different SNRs (5 dB and 10 dB), and processed by 13 different noise reduction algorithms. Using this subjective database [64], Hu and Loizou [44] evaluated several objective speech quality metrics including the PESQ. Indeed, the results from this study showed that the PESQ scores correlated the best with subjective ratings of enhanced

speech quality. This correlation further improved when three other metrics, viz. the LLR, WSS and IS, were combined with the PESQ using the MARS as a cognitive model [44] (described in Chapter 2). Rohdenburg et al. [42] also conducted a similar investigation with a different database. The best objective metrics from their study were PESQ and the Perceptual Similarity Measure (PSM_b) derived using the Perception Model–Quality assessment (PEMO-Q) [53]. It must be noted here that both PESQ and PSM_b computations incorporate auditory models, and were shown to outperform signal-based metrics such as the Signal-to-Noise Ratio (SNR), the IS distance measure, and the LLR [62, 37].

In another study, Salmela and Mattila [45] used a combination of 16 independent objective metrics for evaluating the quality of acoustic noise suppression algorithms. The objective measures they have used were selected from four major categories including: time-domain methods (such as SSNR), spectral distance methods (such as linear or logarithmic spectral distances), perceptual method (such as weighted slope spectral distance (WSSD)) and methods based on linear prediction (IS). These individual objective metrics were combined through linear regression function, computed using the Partial Least Squares Regression (PLSR) technique, which resulted in a high degree of correlation (0.95) with subjective ratings.

Standardization of wideband speech codecs and significant improvements in terms of speech intelligibility and naturalness have increased a thrust in telecommunication industry towards using wider bandwidth speech in all transmission systems; therefore an objective speech quality measure should work well for noise reduction algorithms in wideband context as well.

At present, very few studies have investigated the application of objective speech quality models to wideband noise reduction performance [19, 63, 65]. Moreover, a comprehensive evaluation of noise reduction algorithms themselves is lacking for wideband telephony applications.

This chapter addresses these issues by first developing a subjective quality ratings database for wideband noise reduction algorithms, in a manner similar to the narrowband database developed by Hu and Loizou [64]. In addition, the performance of several auditory-model based objective speech quality measures including LPD, wideband PESQ, HASQI [54] and PEMO-Q [53] was evaluated in terms of predicting the quality of wideband enhanced speech. As discussed in Chapter 2, the performance of these metrics can be further improved by applying a cognitive model on them. As such, the performance of BMLS was evaluated as the "feature mapping" block of objective speech quality estimation. The per-

formance of BMLS was also contrasted with that of MARS [58] in terms of prediction ability and generalization.

The rest of the chapter is organized as follows: Section 3.3 details the noise reduction database and the collection of subjective quality ratings. Section 3.4 presents the results stemming from the statistical analysis of subjective data. The correlations between the objective metrics and subjective quality ratings as well as the performance evaluation of BMLS and MARS are presented in Section 3.5. Section 3.6 details the procedure for fine tuning the noise reduction algorithm performance using the objective metric, and a secondary subjective evaluation of fine-tuned noise reduction algorithm performance. This chapter is then concluded with a summary in Section 3.7.

3.2 Noise Reduction Algorithms

Narrowband noise reduction techniques for telecommunication applications have attracted significant research attention during the past three decades [36]. In general, noise reduction techniques can be categorized into four groups [36]: 1) spectral subtractive algorithms, 2) Wiener filtering algorithms, 3) statistical model-based algorithms, and 4) subspace algorithms.

3.2.1 Spectral Subtractive Algorithms

Spectral subtraction is a simple and relatively effective algorithm in enhancing noisy speech degraded by additive noise. Assuming that the noise is additive and uncorrelated with the speech signal, enhanced speech could be found by subtracting the magnitude (or power) spectrum of the noise from that of the noisy speech spectrum. Let:

$$d(n) = r(n) + v(n) \tag{3.1}$$

where $d(n)$, $r(n)$ and $v(n)$ are noisy speech, clean speech and noise signal, respectively. In the frequency domain, this is written as:

$$D(\omega) = R(\omega) + V(\omega) \tag{3.2}$$

By estimating the magnitude of noise as $|\hat{V}(\omega)|$ using a voice activity detector that identi-

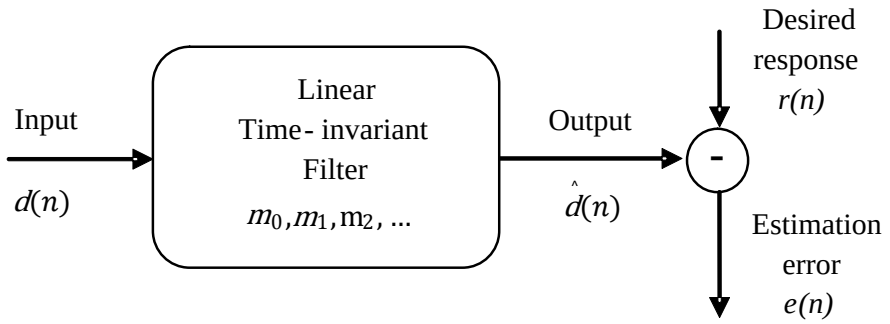

Figure 3.1: Block diagram of the Wiener filter.

fies speech pauses and silence periods, the magnitude of enhanced speech can be computed as $|\hat{R}(\omega)|(=|D(\omega)|-|\hat{V}(\omega)|)$ (this estimate can be extended to power spectrum domain which is involved with estimating the power spectrum of the noise and subtracting it from that of the noisy signal: $|\hat{R}(\omega)|^2(=|D(\omega)|^2-|\hat{V}(\omega)|^2)$). Based on the amount of subtraction, there is a trade off between speech distortion and residual noise of the enhanced speech [36]. Although simple in principle, the estimation of the background noise spectrum is challenging especially in non stationary environments [36]. Multiband Spectral Subtraction (MB) method is an example of this category. This algorithm acts based on the fact that noise will not affect all the spectral components equally and depending on the shape of the noise, some frequencies will be degraded more than the other ones. Therefore the amount of noise reduction, which is determined by a subtraction factor, is different for each frequency band [66].

3.2.2 Wiener Filtering Algorithms

The block diagram of the Wiener filter is shown in Figure 3.1, where $d(n)$ is the noisy speech signal, $r(n)$ is the clean speech, $\hat{d}(n)$ is the enhanced speech, and $e(n)=r(n)-\hat{d}(n)$ is the estimation error. The goal of the algorithm is to compute the filter coefficients m_k, such that the output of the filter is the same as the desired input. This is accomplished through the classical optimum filtering approach, where the filter coefficients that minimize the mean-square estimation error, i.e. $E[e^2(n)]$, are obtained iteratively [36].

It must be noted that the Wiener filter can be derived either in time or frequency domains. However, since the desired clean speech, $r(n)$, is not available in practice, the Wiener filter coefficients must be estimated from the noisy speech alone. Several techniques have been proposed to accomplish this task including the iterative Wiener filter incorporating autoregressive model of speech production, and the constrained Wiener filter that places

both across-time and across-spectral constraints in the deriving the optimal filter [36]. An example of the latter group is the algorithm introduced by Scalart and Filho [67] which uses the decision directed method proposed in [68] for estimating a priory SNR and calculating the filter gain.

3.2.3 Statistical-Model-Based Methods

The main difference between statistical-model-based methods and Wiener filter models is that in the latter the goal is to estimate the complex spectrum of the clean speech while in the statistical methods the focus is on finding a nonlinear estimator of the magnitude of speech spectrum. As an example, if the Discrete Fourier Transform (DFT) coefficients of the noisy speech are available, the goal of the statistical-model-based methods is to estimate the DFT coefficients of the underlying clean speech for resynthesis. Techniques such as maximum likelihood estimation and Bayesian estimation fall under this category of noise reduction algorithms [36].

Minimum Mean Square Error (MMSE) and logMMSE are examples of the algorithms working based on maximum likelihood estimation. These algorithms estimate the magnitude spectra and log magnitude spectra, respectively, by minimizing the mean square error. Another algorithm of this group is log-MMSE-SPU which is based on the fact that speech may not be present at all time and there are some pause periods even during speech activity. This Speech Presence Uncertainty (SPU) is taken into account by involving a factor which shows the probability of the presence of the speech at a particular frequency [36].

WCosh and IS measures are examples of Bayesian estimator categories. Cosh measure is a symmetric distortion and has been derived as a combination of two forms of the IS distortion measures; there is a weighted version of Cosh measure (WCosh) which reflect the auditory masking effects [36].

3.2.4 Subspace Algorithms

Subspace algorithms undertake a different approach to noise reduction and their principle of operation is rooted in the linear algebra theory. The idea behind subspace methods is to decompose the vector space of the noisy speech into two subspaces: the clean speech subspace and the noise subspace. After the successful subspace decomposition, the clean signal can be estimated by nulling the components of the noisy signal in the noise subspace.

These methods try to minimise the speech distortion while keeping the residual noise under a preset threshold value. There are different methods for decomposing the signal into two subspaces [36], of which the Karhunen- Loeve Transform (*KLT*) is most popularly employed subspace decomposition scheme [69].

3.3 Subjective Database for Wideband Noise Reduction

3.3.1 Noise Reduction Algorithms

In this work the performance of six noise reduction algorithms were evaluated including three statistical-model-based algorithms (termed *logMMSE*, *logMMSE_SPU*, *WCosh*), one spectral subtraction algorithm (termed *MB*), one Wiener filtering algorithm (termed *Wiener_as*) and one subspace approach algorithm (termed *KLT*). The MATLAB-code in Loizou [36] were used for implementing all the algorithms. Table 3.1 provides a summary of the algorithm parameters used for implementing the wideband versions, along with the references for the algorithm description and implementation. In addition to the wideband algorithms, the narrowband version of the *logMMSE* algorithm was also included in subjective tests, as this algorithm performed the best in Hu and Loizou's narrowband noise reduction study [64].

3.3.2 Database

A custom database was created for this work based on 16 clean speech samples produced by two male and two female speakers. The clean speech samples were taken from the TSP speech database with an original sampling rate of 48 kHz [72], which were subsequently down-sampled to 16 kHz for our application. The list of the sentences is given in Table 3.2.

Each sentence was corrupted by 3 types of noise (babble, traffic and white) at three SNR levels (0 dB, 5dB and 15 dB). The noisy speech samples were processed by seven aforementioned algorithms (six wideband algorithms and one narrowband algorithm). Thus, the database contained 144 noisy sentences (16 sentences $\times$ 3 noise types $\times$ 3 SNR levels). The total of 144 conditions was partitioned into four parts such that each part included speech samples from four talkers in all the conditions (noise types, SNR levels and all the noise reduction algorithms).

Table 3.1: List of the algorithms used in creating the wideband noise reduction database and the values of the algorithm parameters.

Algorithm	**Parameters**	**Reference**
logMMSE	Eq. 7: $\alpha = 0.99$ Eq. 8: $\beta = 0.99$	[70] [64]
*logMMSE*_SPU	Eq. 7: $\alpha = 0.92$ Eq. 8: $\beta = 0.9$	[70] [64]
Weighted Cosh (WCosh)	Eq. 7: $\alpha = 0.98$ Eq. 34: $p = 1$	[70] [71]
Multiband (*MB*)	Eq. 4.5: $\alpha = 0.9$ Smoothing factor: $\beta = 0.9$	[66]
Wiener_as	Eq. 4: $\eta = 0.15$ Eq. 7: $\beta = 0.98$ Eq. 4: $\lambda = 0.98$	[70] [70] [67]
KLT	VAD parameter: 0.98 VAD threshold: 1.2 Eq. 48: $\mu_0 = 8.2, S = 3.125$ Eq. 48: $a = 2, b = 10$	[36] [36] [69] [69]

Table 3.2: List of sentences used for the wideband noise reduction database.

Filename	**Speaker**	**Gender**	**Sentence text**
Sp01.wav	MA	M	The birch canoe slid on the smooth planks.
Sp02.wav	MA	M	Her purse was full of useless trash.
Sp03.wav	MA	M	Read verse out loud for pleasure.
Sp04.wav	MA	M	Wipe the grease off his dirty face.
Sp05.wav	MJ	M	Clams are small, round, soft and tasty.
Sp06.wav	MJ	M	The line where the edges join was clean.
Sp07.wav	MJ	M	A white silk jacket goes with any shoes.
Sp08.wav	MJ	M	Stop whistling and watch the boys march.
Sp09.wav	FD	F	She has a smart way of wearing clothes.
Sp10.wav	FD	F	Bring your best compass to the third class.
Sp11.wav	FD	F	The club rented the rink for the fifth night.
Sp12.wav	FD	F	Jazz and swing fans like fast music.
Sp13.wav	FG	F	He wrote down a long list of items.
Sp14.wav	FG	F	The drop of the rain made a pleasant sound.
Sp15.wav	FG	F	Smoke poured out of every crack.
Sp16.wav	FG	F	The desk was firm on the shaky floor.

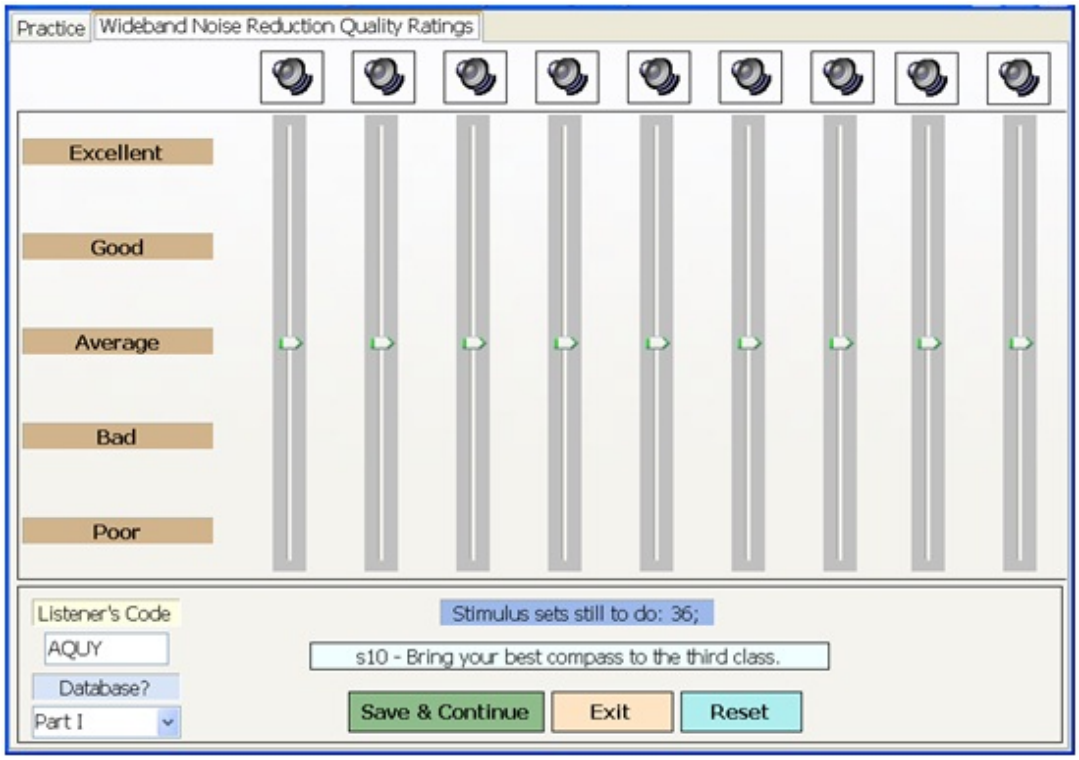

Figure 3.2: Screenshot of the MUSHRA quality ratings software.

3.3.3 Subjective Data Collection

Participants and Audio Presentation

Normal hearing listeners were recruited and each listener rated the quality of all speech stimuli in one of the database parts using the procedure described below. The participants were young adults (16 men, 16 women) at the Western University. Of the participants, 19 were Audiology students, 7 were Speech Language Pathology students, 3 were engineering students, and 3 were not students. None of the participants had explicit knowledge of the stimuli beforehand. English was the first language for all participants. Hearing thresholds were measured using insert transducers at audiometric frequencies between 250-8000 Hz and found to fall in the "normal" range (below 25 dB HL) at each test frequency. For the sound quality ratings task, participants were seated in a double-walled sound booth and listened to stimuli over Sennheiser HDA 200 headphones.

Test Methodology

MUSHRA (MUltiple Stimulus test with Hidden Reference and Anchors) software was used for this experiment [34] (see screenshot in Figure 3.2) and the sound level was adjusted to

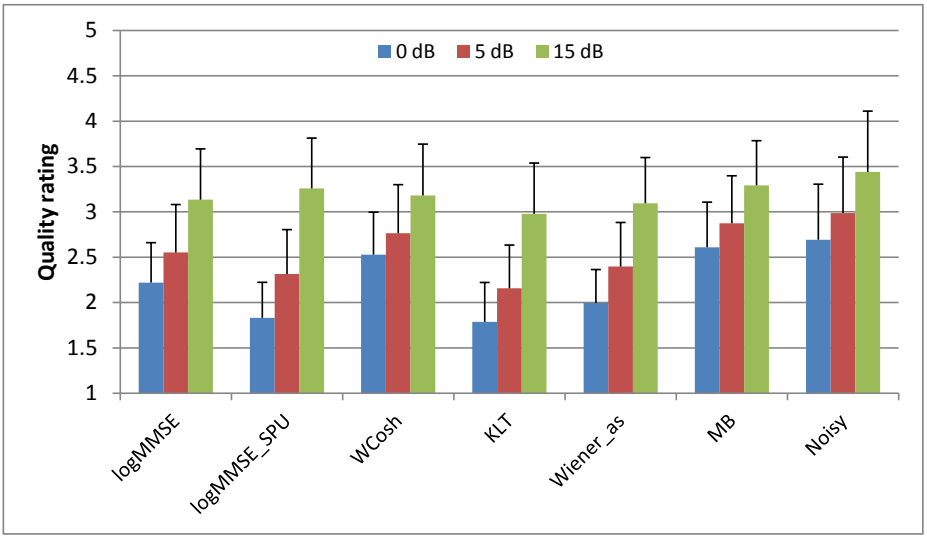

Figure 3.3: Speech quality ratings for the multi-talker babble condition.

be at a comfortable listening level for the participant. There were 36 different screens that participants worked through and each screen had nine speaker icons. These icons were randomly associated with a clean speech sample, its noisy version, and the enhanced version by seven different noise reduction algorithms. Participants were told that they would hear sentences in background noise and they could listen to each sentence as many times as they liked by clicking on the speaker icons. The textual content of each sentence was displayed at the bottom of the screen so that participants know the conveyed message. Participants were instructed to rate each stimulus using the sliders, by paying particular attention to the clarity, pleasantness, distortion/artefacts, and their overall impression of sound quality. When they were satisfied with the rating for each speaker, participants were instructed to click "Save and Continue" button to get to the next screen. To ensure that fatigue was not a confound, this experiment lasted about an hour and participants were encouraged to take breaks when they felt fatigued.

3.4 Subjective Score Analysis

3.4.1 Reliability of the Ratings

The quality ratings were first analyzed for the inter- and intra-rater reliability. Ten of the 32 participants were asked to come back at a later date and redo the rating task. The test-

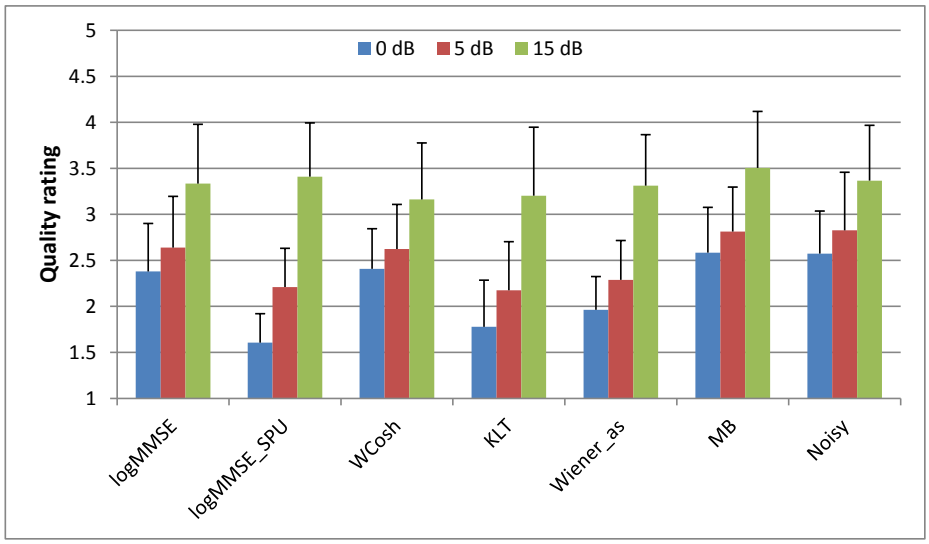

Figure 3.4: Speech quality ratings for the traffic noise condition.

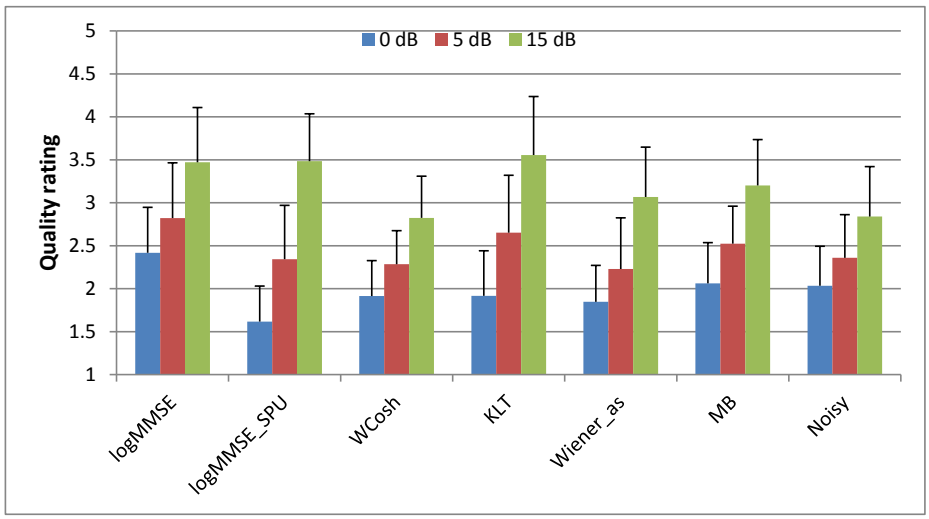

Figure 3.5: Speech quality ratings for the white noise condition.

retest scores for these 10 participants exhibited correlation coefficients ranging between 0.77 to 0.93, indicating high degree of intra-rater reliability. The consistency of ratings among the 32 participants was measured using Cronbach's α and Intraclass Correlation Coefficient (ICC) using SPSS statistical software package, version 19.0. For this dataset, the Cronbach's α was 0.98 and the average measures ICC ranged between 0.924–0.973, once again suggesting a very good agreement among listeners on the quality scores.

3.4.2 Averaged Ratings

Figure 3.3 – Figure 3.5 depict the averaged speech quality ratings for babble, traffic, and white noise respectively. In these figures, the error bars represent one standard deviation. Figure 3.3 – Figure 3.5 show that the speech quality ratings improve with increased SNR. Furthermore, inter-algorithm differences are apparent in these figures across all SNRs and noise-types. In order to quantify the significance of these differences, a thorough statistical analysis was performed.

3.4.3 Statistical Analysis

A split-plot repeated measures analysis of variance (ANOVA) was first performed on the speech quality scores for wideband noise reduction algorithms with noise type, SNR, and algorithm as the within subject factors. Significant two-way interactions were found between noise and SNR ($F(4,28)=4.684$, $p < 0.05$), noise and algorithm ($F(14, 18) = 12.828$, $p < 0.05$), and SNR and algorithm ($F(14,18) = 17.062$, $p < 0.05$). The significant interaction between noise and algorithm variables implies that the algorithm performance depended on the noise type, and the interaction between SNR and algorithm variables implies that the algorithm performance also changed with the SNR.

Using FDR (False Discovery Rate) control method [73], multiple comparisons were performed to assess the significance of the differences between quality scores obtained from different noise reduction algorithms. Tables 3.3 and 3.4 summarize the results. Table 3.3 presents data that will help determine whether there was significant performance difference among algorithms across different noisy conditions, while Table 3.4 displays data that will assist in figuring out whether noise reduction algorithms enhanced quality when compared with the unprocessed, noisy speech sample. Salient features of these two tables include:

Table 3.3: Statistical comparison of different noise reduction algorithms across different noise conditions. "✓" indicates that the algorithms were statistically similar in performance. An absence of "✓" implies that the algorithm performance was inferior.

Noise type	SNR	logMMSE	logMMSE_SPU	WCosh	KLT	Wiener_as	MB
Babble	0			✓			✓
	5			✓			✓
	15		✓	✓			✓
Traffic	0						✓
	5						✓
	15		✓				✓
White	0	✓					
	5	✓			✓		
	15	✓	✓		✓		

Table 3.4: Statistical comparison of ratings of noisy speech and its enhanced version by different noise reduction algorithms. "✓" indicates statistically significant enhancement of speech quality. "ns" indicates that there was no significant difference, while a blank cell implies that the quality degraded after processing.

Noise type	SNR	logMMSE	logMMSE_SPU	WCosh	KLT	Wiener_as	MB
Babble	0						ns
	5						ns
	15		ns				ns
Traffic	0						ns
	5	ns					ns
	15	ns	ns		ns	ns	ns
White	0	✓		ns	ns		ns
	5	✓	ns	ns	ns	ns	ns
	15	✓	✓	ns	✓	✓	✓

- *MB* and *WCosh* performed consistently in babble noise. While *MB* performed well in traffic noise too, *WCosh*'s performance degraded in that noise condition.

- The interaction between algorithm performance, noise type, and SNR is evident from Table 3.3 as some algorithms performed poorly under certain conditions. For example, *logMMSE_SPU* performed well with the SNR at 15 dB, but was inferior at lower SNRs. Similarly, *KLT* performed well with white noise at higher SNRs, but broke down with other noise types.

- Only *MB* algorithm did not degrade speech quality when compared with the noisy speech sample across all conditions. It also improved the quality at one condition, white noise at 15 dB. All other algorithms degraded the sound quality in at least four noise conditions.

- *logMMSE* improved speech quality in white noise at all SNR levels. There was also a statistically significant improvement in speech quality only in white noise at 15 dB SNR condition, for the *logMMSE_SPU* and *KLT*, *Wiener_as* and *MB* algorithms.

In summary, statistical analyses of the subjective ratings showed that there was no single "winner" among the noise reduction algorithms across all noisy conditions. Rather, several algorithms performed similarly in different conditions. For example, both *MB* and *WCosh* produced equivalent performances for the babble noise conditions, and *KLT* and *logMMSE* performed similarly with white noise conditions at higher SNRs. Only one algorithm, *Wiener_as*, performed significantly worse than all other algorithms across the experimental conditions. Overall, the *MB* algorithm was found to be the most consistent performer across all conditions, with its behaviour inferior to that of *logMMSE* algorithm only in the white noise conditions.

A number of parallels can be drawn between the subjective data reported here and the narrowband noise reduction data reported by Hu and Loizou [64]. For example, Hu and Loizou [64] reported that speech quality enhancement was observed only with a few noise reduction algorithms in a subset of noisy conditions. Similarly, our results revealed that statistically significant enhancement in quality was achieved only with the white noise conditions. While there were improvements in speech quality scores for the *MB* algorithm in traffic noise conditions as well, these were found to be statistically insignificant. There were also contrasting differences between the two datasets. For example, Hu and Loizou [64] found that the *Wiener_as* algorithm performed well at a number of noisy conditions in contrast to our findings. In a similar vein, our data showed that the *logMMSE_SPU* performed as good as other algorithms in a few noisy conditions, while Hu and Loizou [64] reported that the *logMMSE_SPU* was inferior to others in the narrowband application. The differences in these results can be attributed to the noise types and SNRs chosen, and the parameters used in the implementation of the algorithms.

As mentioned earlier, the narrowband version of the *logMMSE* algorithm was included in the database and the rating process to have a direct comparison between the narrowband and wideband algorithms. The subjective data from this algorithm were analyzed separately; not surprisingly, the analyses revealed that the narrowband algorithm was inferior to all of the wideband algorithms. For most of the noise conditions, this algorithm was judged to have degraded the speech quality in comparison to the wideband noisy speech samples.

3.5 Objective Quality Evaluation

The objective model evaluation was performed in two stages; first, the performance of objective metrics including overall LPD, wideband-PESQ, HASQI, PEMO-Q, WSS, IS

and LLR, was evaluated before applying the cognitive model on them.

In the second stage, BMLS and MARS were applied on the LPD feature data. They were also used to combine PESQ, LLR, IS and WSS, to develop a composite measure, in a manner similar to Hu and Loizou[44].

3.5.1 Before Applying the Cognitive Models

LPD was calculated following the procedure described in the previous chapter. The analysis was performed on 32 ms blocks of speech, and each frame was processed by auditory filters with centre frequencies ranged from 3 (26.05 Hz) to 33 (7743 Hz) on the ERB scale. To calculate the loudness pattern with higher precision, an interval of 0.1 ERB was selected which resulted in 301 filters. The dimension of LPD ($X(f_c)$), calculated in Equation 2.10, is 30 which was obtained by averaging the loudness pattern across each ten adjacent filters. BMLS and MARS were used to map $X(f_c)$ into the quality scores. To calculate what is referred as the overall LPD, the loudness pattern distortion per ERB ($X(f_c)$), was summed up across the entire ERB scale.

MATLAB code presented in [36] were used for implementing WSS, IS and LLR; however some changes such as increasing the LPC order in LLR and IS, and increasing the number of frequency bands in WSS in order to cover the bandwidth of wideband speech, were introduced in the computation of these metrics for wideband application. The critical band center frequencies and bandwidths proposed in [26] as well as the ERB-rate were used in computing the WSS (now termed WSS-ERB). Wideband-PESQ was implemented using the software given in [38]. The computation of HASQI and PEMO-Q were based on the MATLAB code provided by Kates [54] and Huber [53] respectively.

Two types of correlation analysis were run on the data; in the first version, objective score was computed for each speech sample and these scores were used for correlation analysis; i.e. all 1008 scores (= 16 $\times$ 3 (types of noise)$\times$ 3 (SNR level) $\times$ 7 (6 algorithms + noisy samples)) were used for evaluating each objective measure.

In the second version, the average score at each condition was used for correlation analysis. In other words, after calculating the objective scores for all speech samples, the scores of all 16 speech samples at a specific condition were averaged (same noise type/level and noise reduction algorithm). These condition-averaged scores were used for correlation analysis. After the averaging process, a total 63 (= 3 (types of noise) $\times$ 3 (SNR level) $\times$

Table 3.5: Estimated correlation coefficient and standard error of estimation for different objective quality metrics for per-sample and condition-averaged analysis.

Measure	**Complete Database**		**Condition-averaged**	
	ρ	σ	ρ	σ
PESQ-WB	0.705	0.448	0.825	0.305
PEMO-Q	0.707	0.447	0.798	0.325
HASQI	0.640	0.485	0.842	0.291
LPD	0.734	0.429	0.82	0.309
LLR	0.555	0.525	0.661	0.405
WSS	0.570	0.519	0.778	0.340
WSS-ERB	0.619	0.496	0.788	0.332
IS	0.242	0.620	0.248	0.532

7 (6 algorithms+ noisy samples)) conditions and consequently 63 pairs of subjective and objective scores were available for condition-averaged analysis.

Table 3.5 displays the correlation coefficient and the standard deviation of the error of the objective metrics using two types of correlation analysis. As can be seen from the table, for the per-sample analysis, LPD scores resulted in the highest correlation ($\rho = 0.73$) with the subjective scores, followed by PEMO-Q and PESQ-WB with the same performance ($\rho = 0.70$). While HASQI's performance was poor for per-sample analysis, it resulted in the highest correlation value for the condition-averaged analysis ($\rho = 0.82$). This was due to a bigger variation among the HASQI scores for speech samples within the same condition, which was removed when averaged across the condition. WSS performance was poor in comparison to the other auditory-model based metrics. However, using the ERB-rate scale for modelling the auditory filterbank bandwidths improved WSS performance and resulted in the higher correlation with the perceptual ratings of quality.

Since the signal-based models do not reflect the signal processing in the peripheral auditory system, it was no surprise that IS and LLR resulted in the lowest correlation ($\rho = 0.24$, $\rho = 0.55$ respectively) when compared with the other models.

Figure 3.6 and Figure 3.7 show scatter plots of PESQ, PEMO-Q, LPD and HASQI scores versus subjective scores, for per-sample analysis and condition-averaged analysis respectively. In these figures, the types of noise and the SNR levels have been coded by different shapes and colors, respectively. As can be seen from the figures, there is a linear relationship between the objective and the subjective scores, with a positive correlation for PESQ, PEMO-Q and HASQI and a negative trend for LPD. The predicted quality scores also tend to form separate clusters for different SNR levels. These clusters were more separate for condition-averaged scores, specifically for LPD and HASQI.

Our results can be compared to previously published data on the effectiveness of these pre-

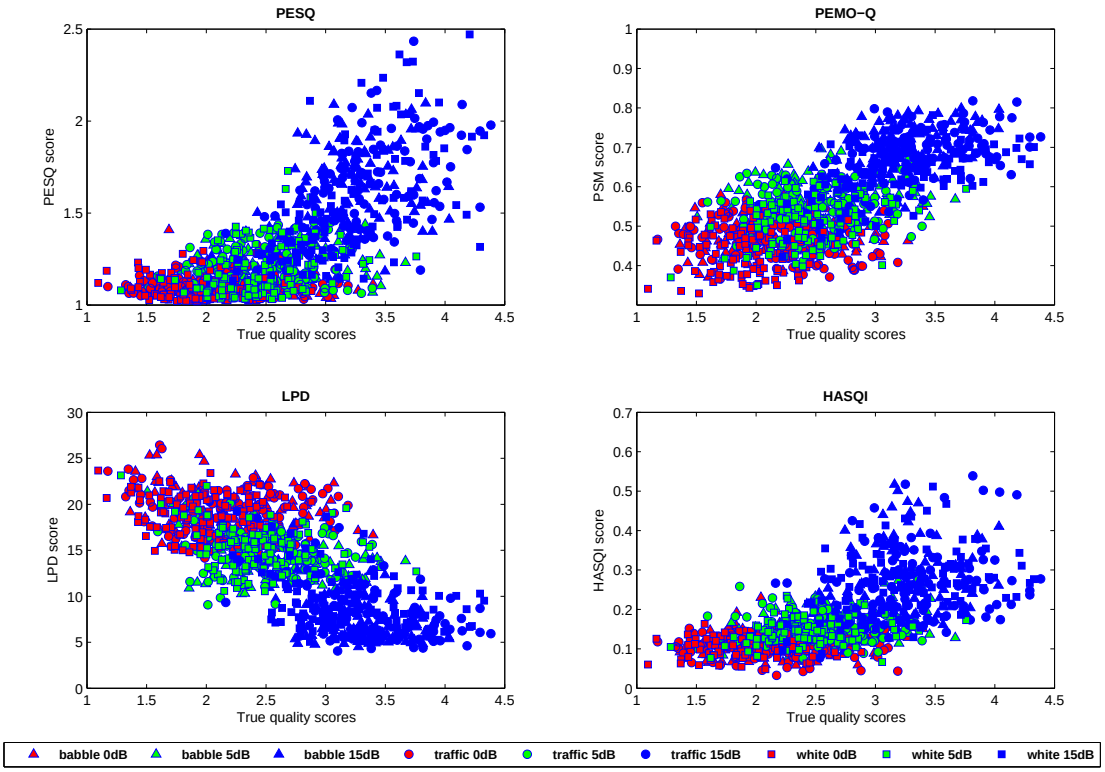

Figure 3.6: Noise and level-dependent relationship between PESQ, PEMO-Q, LPD and HASQI scores versus the true quality scores: per-sample analysis.

dictors. For example, Hu and Loizou [44] showed that PESQ performed the best among all objective metrics in predicting the subjective quality assessments of narrowband noise reduction algorithms. Utilizing the same narrowband database, Kressner *et al.* [43] demonstrated that the performance of HASQI is similar to that of PESQ. Rohdenburg *et al.* [42] evaluated the performance of two versions of the *logMMSE* algorithm through subjective and objective measures, and found that PESQ correlated best with the subjective ratings of post-enhancement speech quality. Our data with a broader set of algorithms and noisy conditions also showed that PESQ performs well in predicting subjective quality ratings.

It is important to note here that even though the correlation values between the subjective scores and the scores predicted by various objective models were different (see Table 3.5), these differences may not always be statistically significant. Thus, it is imperative to test statistical significance of these correlation coefficients, a point not addressed in [44] and [42]. In this paper, Steiger's t-test for dependent correlations with 95% confidence intervals [74] was used to evaluate whether the difference between the correlation coefficients from two different metrics is statistically significant. This test was conducted for

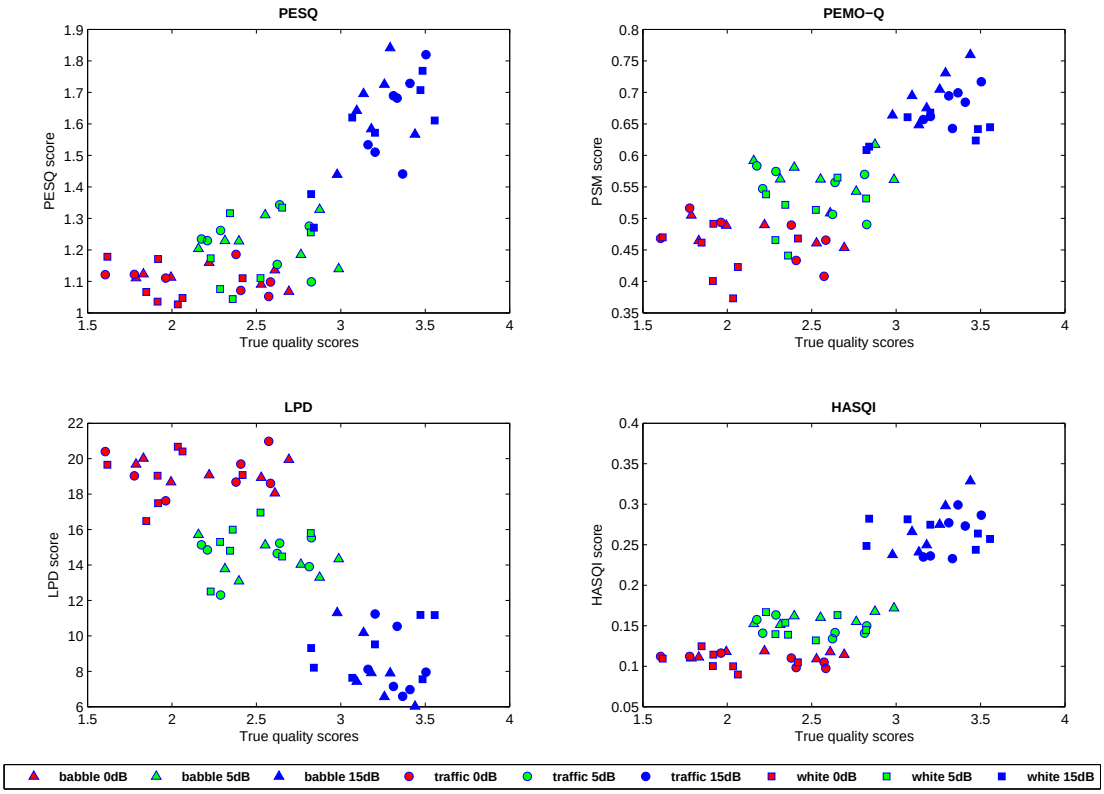

Figure 3.7: Noise and level-dependent relationship between PESQ, PEMO-Q, LPD and HASQI scores versus the true quality scores: condition-averaged analysis.

both raw and condition-averaged correlation coefficients displayed in Table 3.5. When the entire database was considered, Steiger's t-test showed that the LPD correlation coefficient was significantly greater than those obtained for PESQ ($Z = 2.32 > 1.96$), PEMO-Q ($Z = 2.74 > 1.96$) and HASQI ($Z = 7.3 > 1.96$). While there was no significant difference between PESQ and PEMO-Q coefficients ($Z = 0.16 < 1.96$), they were both statistically better than HASQI ($Z = 4.01 > 1.96$, $Z = 4.15 > 1.96$ respectively). With condition-averaged correlation coefficients, however, there were no statistically significant differences among the four auditory model-based metrics.

The scatter plots shown in Figure 3.6 reveal an interesting pattern. While both LPD and PEMO-Q had values scattered around the diagonal, both PESQ and HASQI results were characterized by a "bottoming" effect – the predicted quality scores reached the minimum value, although the true quality scores ranged between 1 and 3. This was due to the regression functions within the PESQ and HASQI computation, which map the feature vector to a predicted quality score. The regression functions were derived from separate databases and the methodological differences in subjective data collection may have rendered the

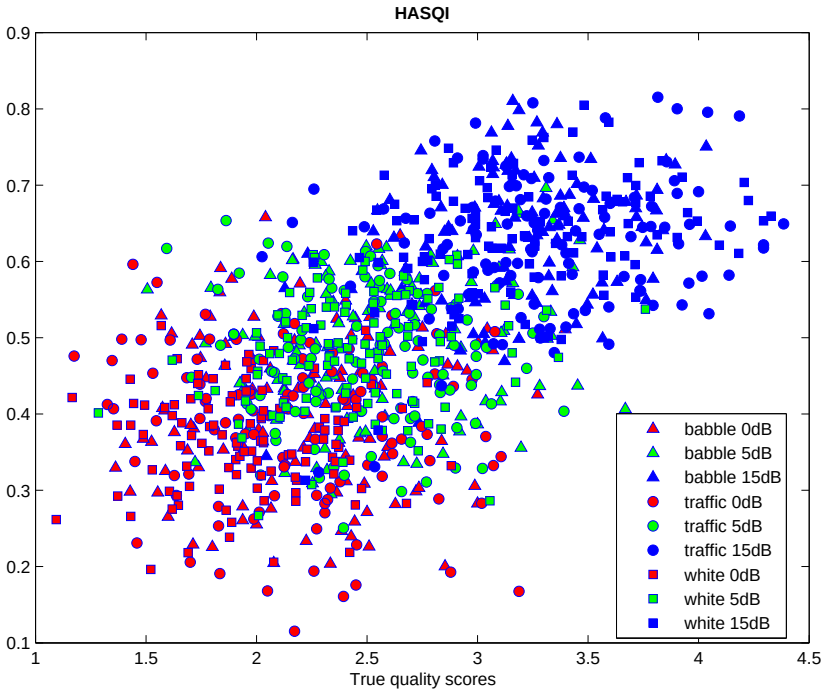

Figure 3.8: Noise and level-dependent relationship between cepstrum correlation (CC) value (generated by the HASQI procedure) versus the true quality scores: per-sample analysis.

mapping process suboptimal for our database. This hypothesis was tested by analyzing the raw feature from HASQI before the application of the regression function. Figure 3.8 shows the scatter plot between the true quality score and the cepstrum correlation (CC) value generated by the HASQI procedure. It is evident that this scatter plot resembles those from LPD and PEMO-Q. This finding highlights the role of regression functions within the speech quality estimators and their ability to generalize across different databases.

Computational models of human audition continue to be proposed and improved upon. For example, PEMO-Q's auditory model is based on the model proposed by Dau et al. [51, 52]; the model has been successful in a variety of applications such as speech intelligibility prediction [75], speech recognition [76], speech quality evaluation [53], and also featured in our study. Jespen *et al.* [50] revised this model by modifying peripheral and central stages of the model. The modified model, called the computational auditory signal-processing and perception (CASP), substituted the gammatone filters with a dual resonance nonlinear (DRNL) filterbank, which better models the nonlinear characteristics of the basilar membrane such as level- and frequency-dependent compression. The other major modification was the addition of an "expansion" block before the adaptation process, which models the square law behaviour of rate-versus-level function of auditory nerve systems. Using

our database, the performance of this enhanced model in predicting speech quality scores was investigated. Results showed that the new model did not enhance the performance of PEMO-Q. This suggested that while the new CASP model may better explain psychoacoustic phenomenon of intensity discrimination with pure tones and broadband noise, spectral masking with narrowband signals and maskers, and amplitude modulation detection with narrow and wideband noise carriers, it does not extract new information to aid in speech quality estimation.

3.5.2 After Applying the Cognitive Models

There are two basic steps for evaluating the performance of each cognitive model:

1- Model selection: which involves optimizing the model parameters and choosing a good parameter set for the model.

2- Performance evaluation: which deals with comparing the performance of the model with the other models in terms of predictive ability and generalization.

Model Selection

As mentioned before, in Bayesian analysis the number of basis functions as well as the position of them are determined by the data on which the model is trained and therefore the complexity of the model is determined with the training data. The only parameters which should be set for BMLS are the prior probability parameters, which include: a and b (parameters of *InvGamma* distribution of σ^2), Z (maximum interaction level) and λ (the precision of the normal distribution on β).

Many different parameter sets were selected with $a, b \in [0,1]$, $Z \in [3, dim(X)]$ and $\lambda \in \{1, 0.1, 0.01, 0.001, 0.0001\}$. K-fold cross validation with $K = 4$ was used for choosing the best set; the whole database was split into four sets; each time three sets were used for training the model and the remaining set for testing, and this process was repeated four times. In this way, all the data has been used once for testing the model. Mean square prediction error (MSPE) was used as the cost function:

$$MSPE = \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2 / n \tag{3.3}$$

The best parameter set for a cognitive model depends on the data on which the model is applied. The parameters of BMLS for LPD, and the combination of the other four objective

Table 3.6: The effect of change in the "maximum allowed interaction level" parameter

Max interaction level	MSPE
2	0.1587
10	0.1527
15	0.1526
20	0.1506
25	0.1514
30	0.1503

Table 3.7: The effect of change in the "precision" parameter

Precision	MSPE
1	0.1521
0.1	0.1522
0.01	0.1559
0.001	0.1617
0.0001	0.1636

metrics were optimized separately.

The averaged MSPE for several values of Z, λ and (a,b) for LPD data, is shown in Tables 3.6, 3.7 and 3.8, respectively. It should be mentioned that in each table, except the parameter that was varied, the values of the other parameters were fixed.

As can be seen from the table, there is no significant changes in MSPE value with changes in the parameters; it shows that these parameters does not play critical role in the model's performance. The only parameter which has a slight effect on the MSPE value is λ parameter, which is the prior parameter for the inverse variance of the basis functions weights. This parameter affects the number of basis functions required for the regression model. Figure 3.9 shows the histograms of the posterior samples of parameter k for $\lambda = 1, \lambda = 0.1, \lambda = 0.01$ and $\lambda = 0.001$. From Figure 3.9, it can be observed that for large $\lambda(\lambda = 1)$, since the prior parameter does not give much flexibility to the basis functions, a

Table 3.8: The effect of change in the parameters of *Invgamma* distribution

(a,b)	MSPE
$(1,1)$	0.1552
$(0.01,0.1)$	0.1564
$(0.01,0.01)$	0.1557
$(0.001,0.001)$	0.1532
$(1,0.001)$	0.1568
$(0.001,1)$	0.1553
$(0,0)$	0.1572

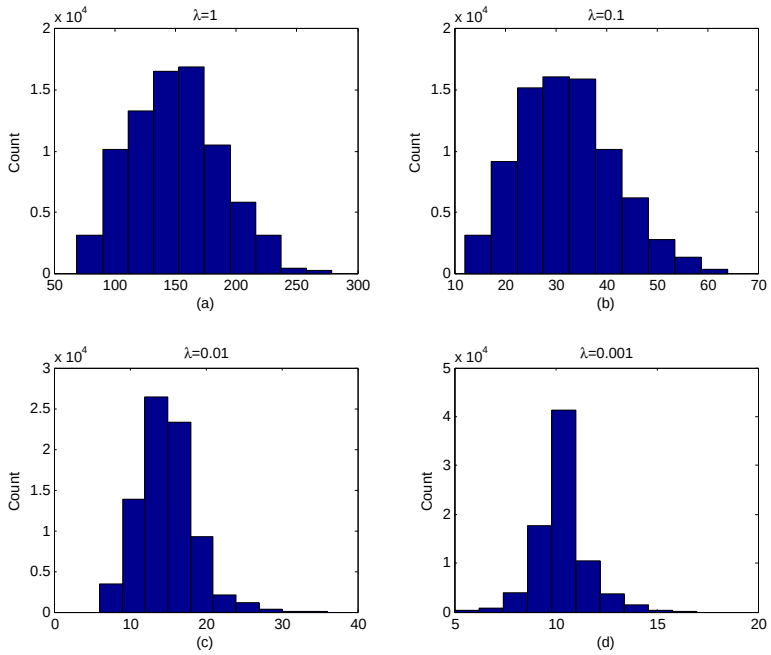

Figure 3.9: Histograms of the posterior samples of k for (a)λ=1;(b)λ=0.1;(c)λ=0.001;(d)λ=0.0001

larger number of basis functions are required to model the regression function. In contrast, small $\lambda(\lambda = 0.0001)$ results in fewer basis functions, since each basis function is more flexible and has more degrees of freedom.

Finally, the parameters a,b, λ and Z were fixed at 0.001, 0.001, 0.1 and 10 respectively, for LPD data. The model parameters were also optimized for combining PESQ, WSS, IS and LLR following the same procedure described above, and same as before, the parameters did not have a significant effect on the model performance. Except for the Z parameter which was set to 4, the other parameter values were set to the same values used for LPD data.

The first 60000 generated samples were discarded as a burn-in period and every fifth of the next 80000 models was used for the analysis.

The 4-fold cross validation procedure was also followed for optimizing the parameters of MARS for LPD data as well as the combination of PESQ, WSS, IS and LLR. Maximum number of basis functions and the interaction level were the two main parameters of MARS model which should be set before using the model. These two parameters were respectively set to 30 and 10 for LPD data, and 30 and 4 for the composite measure.

Performance evaluation

After finalizing the parameters for MARS and BMLS, the performance of the models was

compared using 4-fold cross validation. The averaged value of the correlation coefficients was reported as the final correlation value. Table 3.9 shows the results.

It can be seen from the table that the combination of four objective measures (PESQ, LLR, IS and WSS) using BMLS resulted in the highest correlation among the other metrics (ρ=0.81 for per-sample analysis and ρ=0.91 for condition-average one). Combination of these metrics using MARS, yielded the second top model (ρ=0.79 for per-sample analysis and ρ=0.91 for condition-average one). BMLS-LPD stood as the third best method (ρ=0.78 for per-sample analysis and ρ=0.86 for condition-average).

Steiger's Z-test for dependent correlations with 95% confidence intervals [74] was used to indicate whether the differences between the correlation values were statistically significant or not. The results showed that there were statistically significant differences between LPD-BMLS and PESQ-WSS-IS-LLR-BMLS for both per-sample $(Z = 2.99, p < 0.05)$ and condition average $(Z = 2.97, p < 0.05)$ analysis, as well as between LPD-MARS and LPD-BMLS for per-sample analysis $(Z = 4.05, p < 0.05)$. There was no statistically significant difference between (PESQ, LLR, IS and WSS)+MARS and (PESQ, LLR, IS and WSS)+BMLS neither for per-sample analysis $(Z = -1.86, p > 0.05)$ nor for condition-average one $(Z = 0.06, p > 0.05)$.

Although LPD-BMLS method resulted in lower correlation than the other two composite measures (combination of PESQ, LLR, IS and WSS using MARS and BMLS), it should be noted that implementing LPD-BMLS is computationally simpler than BMLS-(PESQ, LLR, IS and WSS). For implementing the former one, the loudness pattern coefficients should be calculated, which involves modelling the peripheral auditory system, while the second one involves calculating the PESQ score, which includes models of the peripheral auditory system, as well as three more metrics, LLR, IS, WSS, where the first two are LPC-based metrics and WSS is based on modeling the auditory filterbank and calculating the weighted spectral slopes in each band.

Figure 3.10 depicts the scatter plot of the predicted values of the overall quality scores versus the actual scores for LPD coefficients before and after applying BMLS.

The performance of BMLS and MARS was similar when they were used for mapping the combination of four objective metrics into the quality score. A comparison between LPD-MARS and LPD-BMLS showed that BMLS improved the performance of LPD more; since the dimension of LPD coefficients is 30 (30 auditory filters were used for calculating loudness pattern), these results confirmed the fact that MARS is suitable for data with moderate dimensions $3 < N < 20$ [58] and BMLS is an alternative for MARS for higher

Table 3.9: Estimated correlation coefficient ρ and standard error of estimation (σ_e) for various objective quality metrics after applying mapping function.

Objective measure	**Complete Database**		**Condition-averaged**	
	ρ	σ	ρ	σ
LPD+BMLS (testing)	0.7821	0.3921	0.8555	0.2890
LPD+BMLS (training)	0.8381	0.3435	0.9114	0.2221
LPD+MARS (testing)	0.7529	0.4156	0.8377	0.3034
LPD+MARS (training)	0.8027	0.3767	0.8957	0.2409
PESQ+LLR+IS+WSS+BMLS (testing)	0.8079	0.3715	0.9103	0.2310
PESQ+LLR+IS+WSS+BMLS (training)	0.8247	0.3570	0.9488	0.1710
PESQ+LLR+IS+WSS+MARS (testing)	0.7936	0.3842	0.9117	0.2295
PESQ+LLR+IS+WSS+MARS (training)	0.8310	0.3512	0.9571	0.1571

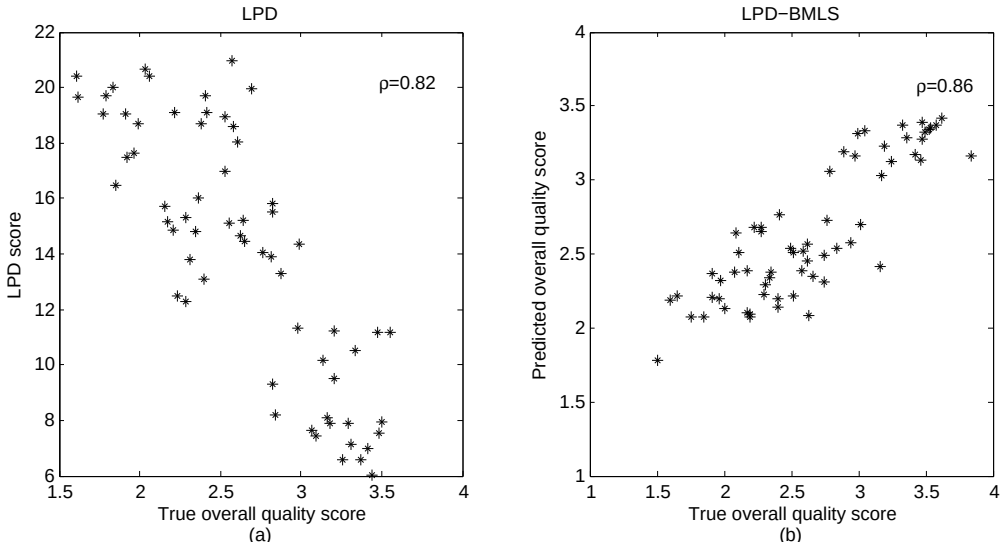

Figure 3.10: Scatter plot of the condition-average predicted and actual overall quality scores for (a) LPD coefficients ($\rho = 0.82$) and (b) LPD-BMLS scores ($\rho = 0.86$).

dimensional regression problems [59, 56].

3.6 Fine-tuning the Noise Reduction Algorithm

As shown in the previous section, quality scores predicted by LPD-BMLS are highly correlated with the subjective scores. This high degree of correlation can be exploited for updating the parameters of noise reduction algorithms such that their performance is optimized for wideband applications. This is based on the premise that if the objective metric was validated, then the algorithm parameter set which maximizes the metric will maximize the subjective quality score as well; the higher the quality score predicted by the model, the

higher the quality of the processed speech sample. However, after updating the algorithms, another subjective test is required to verify this hypothesis.

The parameters of the noise reduction algorithms used in the earlier section were fine-tuned using the LPD-BMLS metric. The performance of the algorithms before and after the updates were also compared together. These processes are explained in more details in the following.

3.6.1 Training the Model

The customized database and subjective quality scores were used for training the model. After the training phase, the parameters of BMLS basis functions and their weight coefficients were stored to be used in the next steps for updating the parameters of the noise reduction algorithms.

3.6.2 Updating the Algorithm

For each algorithm, those parameters were selected which could have significant effect on the quality of speech with the bandwidth increase. Some of these parameters included smoothing factors, threshold values for VAD detectors, threshold values for noise estimation, and FFT length. The effects of FFT length and the parameter of first-order recursive filter have been shown in [20].

Using noise reduction algorithms with the specified sets of parameters, the noisy speech samples were enhanced and the objective quality scores were calculated for them. Since the speech samples were corrupted by three types of noise at three SNR levels, for each condition the parameter set which resulted in the maximum objective score was selected; therefore there were nine (three noise types $\times$ three SNR levels) best parameter sets for each noise reduction algorithm. To select the set which performed well across all the noisy conditions, the outputs of the algorithm fine-tuned by all nine parameter sets were carefully listened to, and the set which resulted in the best quality across all conditions was identified. This set was used for implementing the algorithm for the wideband application. The same procedure was repeated for all the algorithms.

Table 3.10 provides a summary of the changes made to the algorithm parameters, based on the objective quality scores. The number in the parenthesis show the values of the

Table 3.10: List of algorithms used in creating the updated wideband noise reduction database, and the values of the algorithm parameters.

Algorithm	Noise Type	Changed parameters	Reference
logMMSE	Traffic 15dB	Eq.: 4 η=0.15 (0.15) Eq. 7: α=0.95 (0.99) Eq. 8: β=0.97 (0.99)	[70] [70] [64]
MB	Babble 5dB	Eq. 4.5: α=0.96 Smoothing factor: β=0.97	[66] [66]
KLT	White 0dB	VAD parameter: 0.98 (0.98) VAD Threshold: 1.4 (1.2) Eq. 48: μ_0= 4.2(8.4), s=6.25 (3.125)	[36] [36] [69]
Wiener_as	White 15dB	Eq. 4: η=0.6 (0.15) Eq. 7: β=0.99 (0.98) Eq. 4: λ=0.94 (0.98)	[70] [70] [67]

parameters for making the first database. The table also displays the noise condition that resulted in the selected parameter set.

3.6.3 Verification of the Model Validation

Since subjective evaluation is a time-consuming and expensive task, a second subjective speech quality rating experiment was conducted only for two of the algorithms,"*MB*" (spectral subtraction class) and "*logMMSE*" (statistical class).

A new database was created using the procedure described in Section 3.3.2. The database contained 16 speech samples corrupted by three noise types at three SNR levels and enhanced by two noise reduction algorithms. A total of 29 subjects rated the quality of the database.

The quality ratings were first analyzed for the inter-rater reliability. The Cronbach's α was 0.98 and the average measures ICC ranged between 0.92 - 0.98; These numbers suggest a very good agreement among listeners on the quality scores, very similar to the reliability data obtained in the first experiment.

The correlation between the subjective quality scores of the updated algorithms and their LPD-BMLS predicted scores was calculated; the correlation between the objective and subjective scores averaged across similar conditions in the entire database was calculated as well. The correlation values were 0.79 and 0.91 respectively, which further validated the performance of LPD-BMLS. Figure 3.11 depicts the scatter plot of the condition-average predicted values of the overall quality versus the condition-average subjective quality scores for updated "*MB*" and "*logMMSE*" algorithms.

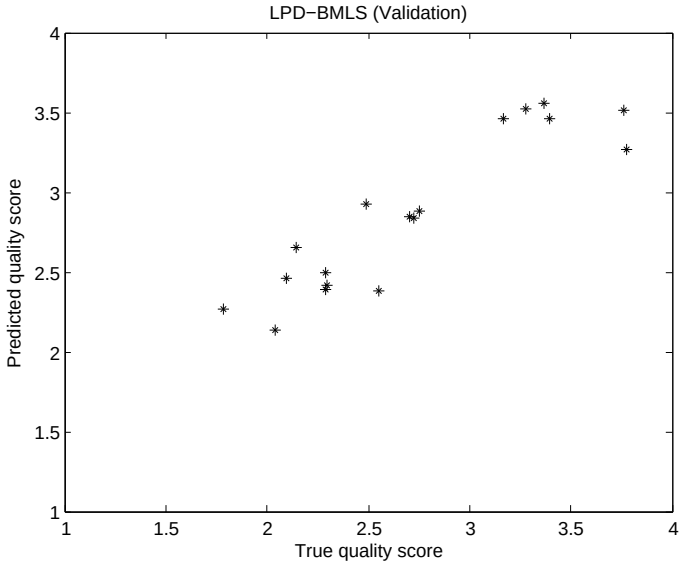

Figure 3.11: Scatter plot of the BMLS-LPD condition-average predicted scores vs. actual overall quality scores of the updated algorithms ("*MB*" and "*logMMSE*")

3.6.4 Paired Comparison Between the Algorithms Before and After the Updates

To have a direct comparison between the performance of the algorithms before and after the updates, another database containing speech samples enhanced by both new and old algorithms was created. The database contained 16 speech samples corrupted by three types of noise at three SNR levels and enhanced by four algorithms, from four different classes of the noise reduction algorithms: "*logMMSE*" (statistical-model-based algorithms), "*MB*"(spectral subtraction), "*KLT*"(subspace approach algorithm) and "*Wiener_as*" (Wiener filtering algorithm).

The experiment was controlled by a panel shown in Figure 3.12. Subjects were asked to compare the quality samples A and B which have been processed by the new and the old noise reduction algorithms, and express their preference ratings by selecting one of the five available options. The presentation order of the algorithms and speech samples were chosen randomly. 17 subjects were recruited for the subjective data collection. Each subject compared 144 pairs (4 speech sample $\times$ 3 types of noise $\times$ 3 SNR levels $\times$ 4 noise reduction algorithms).

For data analysis, the preference ratings were mapped to a five-point scale: $\gamma_{1,2} = -2,-1,0,1,2$ and $\gamma_{2,1} = -\gamma_{1,2}$ which captured the degree of preference for the new algorithm over the old one, and vice versa respectively.

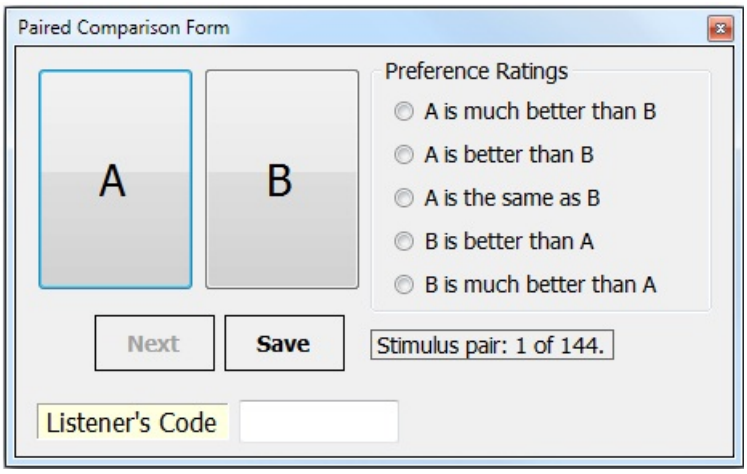

Figure 3.12: Screenshot of the software used for paired comparison of the algorithms

Since the subjects were not instructed on how to make the judgments using the whole scale of rating, it cannot be assumed that the five answer categories were equivalent across different subjects; therefore the assigned numbers to the five options do not necessarily reflect the perceived differences between the signals and the subjects. These numbers were therefore transferred to preference scores using the method presented in [77].

Hansen [77] calculated two different types of scores. For each algorithm, the "simple preference score" shows how often the new algorithm was preferred over the old algorithms (δ_1), and vice versa (δ_2), but it does not give any information about the strength of the preference:

$$\delta_i = \sum_{j,k,(\gamma_{ij}(k)>0)} \text{sign}(\gamma_{ij}(k)), \quad i = 1,2 \tag{3.4}$$

where $\gamma_{ij}(k)$ shows the assigned number for speech sample k, and $sign(\gamma_{ij}(k))$ is equal to one for positive $\gamma_{ij}(k)$ and 0 otherwise.

The "weighted preference score(ω)" shows how often and with what strength the subjects preferred the new algorithm over the old one (ω_1), and vice versa (ω_2).

$$\omega_i = \eta \sum_{j,k,(\gamma_{ij}(k)\geq 0)} \gamma_{ij}(k), i = 1,2 \tag{3.5}$$

which η is a normalization constant and is calculated for each subject individually to equalize the influence of the answering scale usage by the subjects. η is selected so that:

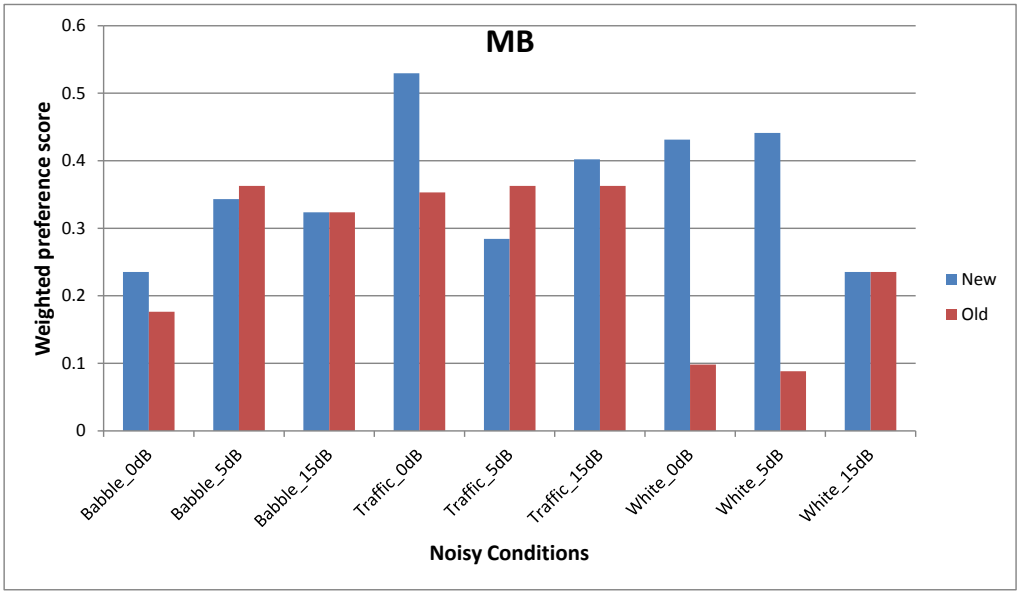

Figure 3.13: Preference ratings of the algorithms after fine-tuning by the objective model - *MB*

$$\sum_{i=1}^{2} \omega_i = 1 \tag{3.6}$$

Figures 3.13 - 3.16 depict the weighted preference scores (ω_i) averaged across all subjects, for the new and the old algorithms at different noisy conditions for *MB*, *logMMSE*, *Wiener_as* and *KLT* algorithms respectively.

For the statistical data analysis, for each algorithm and at each noisy condition, a Wilcoxon signed tanks test was first applied on the simple preference scores to see if there were statistically significant differences in the preferences for the new and old algorithm. If any significant differences were observed, then the same test was also performed on the weighted preference scores. Table 3.11 summarizes the results of these tests.

The table shows that the updated "*logMMSE*" resulted in the most improvements by performing better at several noisy conditions such as 5 dB and 15 dB traffic noise, and 0dB and 5 dB white noise. The performance of the fine-tuned "*MB*" was also improved only at low SNR levels white noise and updated "*Wiener_as*" performed better only at 5 dB traffic noise. Except "*KLT*" which had an inferior performance at only one condition, there was no statistically significant differences between the algorithms before and after the updates at most of the conditions.

In summary, the parameter sets selected by the objective methods could only improve the

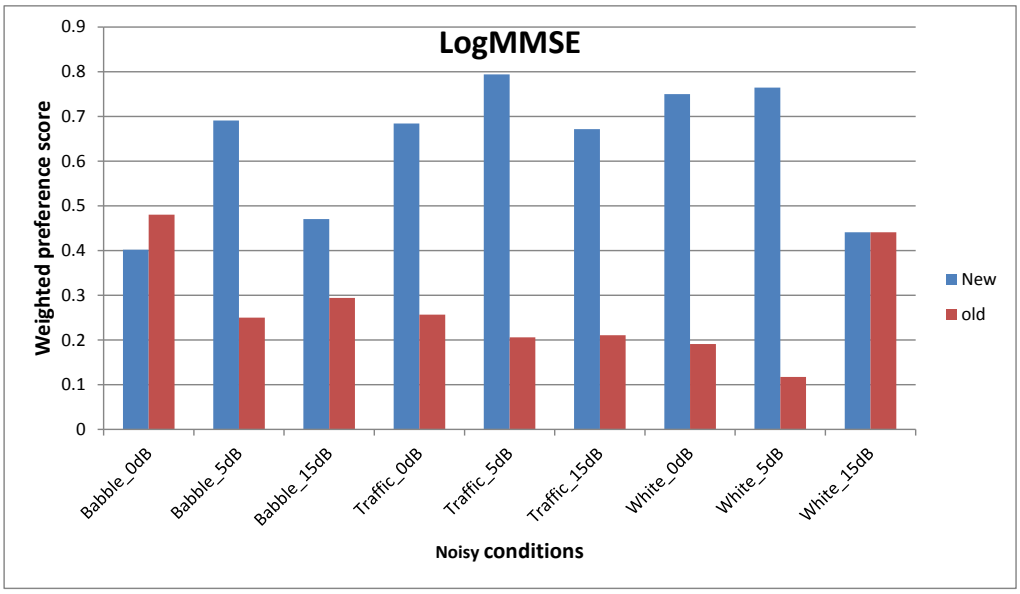

Figure 3.14: Preference ratings of the algorithms after fine-tuning by the objective model - *logMMSE*

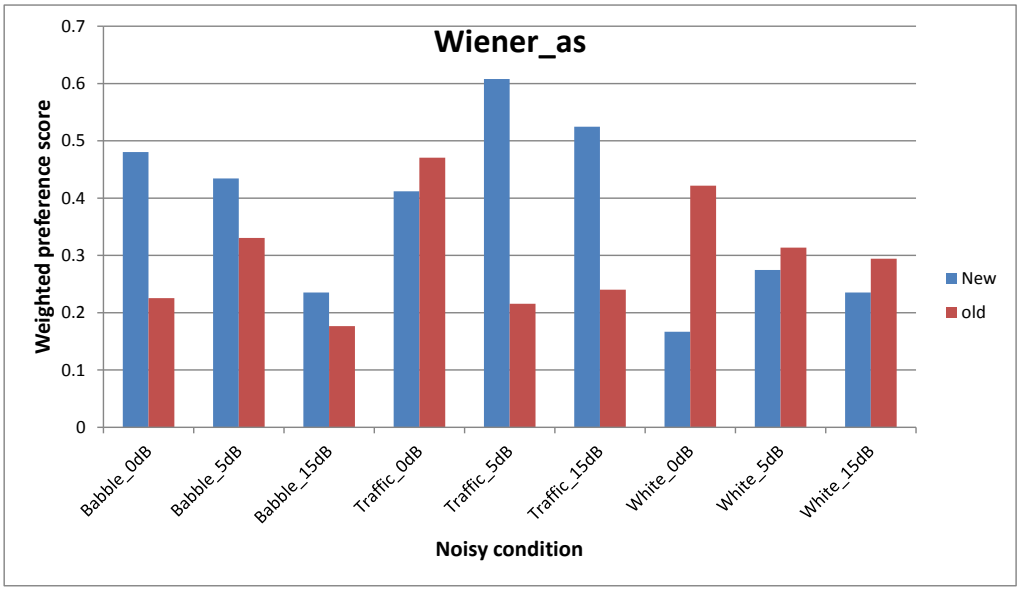

Figure 3.15: Preference ratings of the algorithms after fine-tuning by the objective model -*Wiener_as*

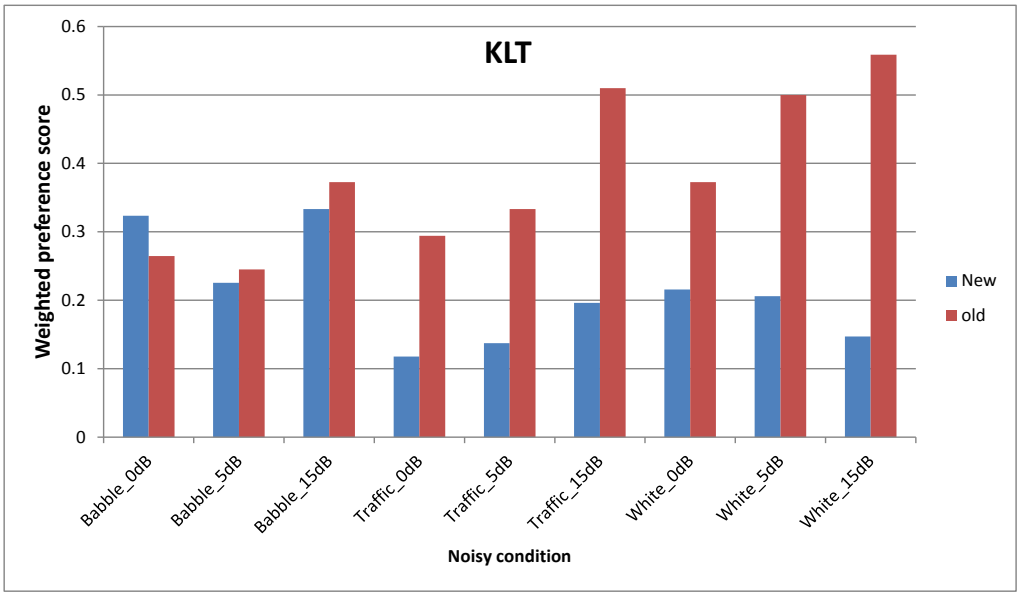

Figure 3.16: Preference ratings of the algorithms after fine-tuning by the objective model - *KLT*

Table 3.11: Comparison of noise reduction algorithms before and after the fine-tuning, across different noise conditions. "+" indicates that the algorithm performance has been improved after the updates; "-" shows that the algorithm performance was inferior after the updates. An absence of "+" and "-" implies that the algorithm before and after the updates, were statistically similar in performance.

Noise type	SNR	*MB*	*logMMSE*	*Wiener_as*	*KLT*
Babble	0				
	5				
	15				
Traffic	0				
	5		+	+	
	15		+		
White	0	+	+		
	5	+	+		
	15				-

performance of some of the algorithms at some isolated noisy conditions. For example there was no change in the algorithms performance in babble noise, before and after the updates. These results show that there is no unique parameter set which would improve the algorithms performance in all types of noise and at all SNR levels. The same conclusion can be drawn from the study conducted by Hu and Loizou [64] where the performance of these algorithms, optimized for narrowband application, were evaluated and compared and none of them were shown to improve the speech quality across all noisy conditions, in comparison with the noisy signal.

3.7 Summary

The present chapter reported the performance of auditory model-based speech quality prediction methods. The prediction accuracy of the ITU-standardized PESQ, a metric based on Moore-Glasberg loudness model (LPD), a metric based on Dau et al., model (PEMO-Q), and the recently proposed HASQI, was compared. To the best of our knowledge, data comparing the performance of all of these auditory models is lacking, especially for the assessment of noise reduction algorithms in wideband context.

The performance of BMLS in predicting the quality of wideband noise-reduced signal was also evaluated and compared with MARS. The models were used to map the LPD feature vectors into the quality scores. They also were employed in combining four objective metrics: PESQ, WSS, IS and LLR. The models' performance was evaluated and validated using Pearson correlation coefficient and 4-fold cross validation. The results showed that LPD-BMLS had a high correlation with the subjective scores ($\rho = 0.78$ and $\rho = 0.86$ for per-sample and condition-average analysis respectively); however the model did not perform as well as the combination of PESQ, WSS, IS and LLR using BMLS and MARS, but its computationally simpler implementation in comparison with the other, selected the model to be used for fine tuning the parameters of the noise reduction algorithms for wideband application.

Subjective ratings of the algorithms with the new parameters validated the performance of the model. On the four algorithms optimized for wideband application, except "klt" which did not show any improvement, the other ones performances improved at some isolated noisy conditions. None of the algorithms could perform better at all the noisy condition after the updates. Since even in the study on evaluating the performance of narrowband noise reduction algorithms [64], none of the algorithms could improve the quality of speech

at all noisy condition, it seems that there may not be any single parameter set which could make improvements at all the noise types and in different SNR levels.

Chapter 4

Echo Evaluation in Listening Context

4.1 Introduction

Every telephone connection has two transmission paths, one from the first party to the second and another from the second party to the first. Echo is the result of the signal on one path getting into the other path, usually with some delay.

Sources of echo depend on the type of voice terminal in telecommunication systems. There are two main types of terminals: analog and digital. For historical and economic reasons, analog terminals have two wires and carry signal in both directions on the same pair of wires; they are connected to the four-wire central network via a converting device called a hybrid. Because there is always some mismatch in the electrical impedance between the two-wire and four-wire sections of the network, a portion of the signal is reflected back to the sender and causes echo which is called electric echo. In contrast, digital terminals are "four-wire" and as a result there is no hybrid and electric echo associated with digital access links.

Another type of echo is acoustic echo which has its source in acoustical coupling of the send and receive paths. Acoustic echo is generated from the earphone/speaker signal reflected off the objects in the environment and then picked up by the microphone; it is also generated from acoustical coupling inside the terminal or direct path from the speaker to the microphone.

Inductive coupling in the cord of the terminals and electrical/mechanical coupling in the electrical/mechanical circuits of the network are the other sources of echo.

Figure 4.1 shows the voice paths, the delays and the main echo sources in a connection

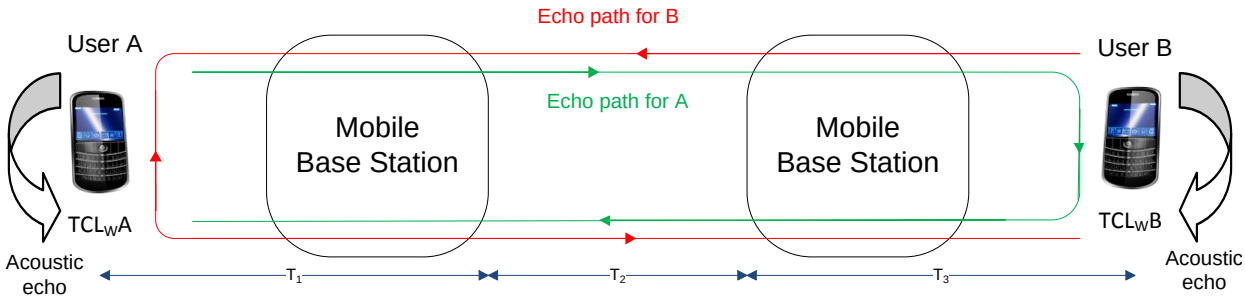

Figure 4.1: Mobile to mobile connection

between two mobile phones. Here, T1 and T3 are delay values resulting from signal processing such as digitalization, compression and coding in the network, and also processing at the radio frequency interface of the network [78, 79].

After introducing satellite communication and increasing intercontinental telephony in 1970s, the effects of delay and echo became more important for the telephone system design engineers [80]. The role of these effects are getting even more critical with the development of modern wireless and packet-based network which introduce more delay to the transmission system [41].

Evaluating the effect of echo and delay can be done through both subjective and objective methods. Subjective evaluation of telecommunication systems can be performed using listening-only or conversational tests. The procedure for doing these tests for different applications have been described in ITU-T P. 800 [27].

While the collection of quality ratings from a group of the listeners is the most reliable method for evaluating the speech quality, it is also time- and resource-consuming. Therefore, an objective metric that correlates highly with subjective data is attractive. As discussed earlier, objective measures can be based on the physical measures of the signals or system under test (parametric models) or input/output speech signals (signal-based models).

Among parametric models, the E-model standardized by ITU-T G. 107 [18], is the most widely used one. This model uses physical measures of the system under test to provide a speech quality score. Some of these parameters are: SLR (Send Loudness Rating), RLR (Receive Loudness Rating), TELR (Talker Echo Loudness Rating), round-trip delay, room-noise (send/receive), circuit-noise and etc. The procedure for deriving equipment impairment factors from subjective listening – only test has been described in ITU-T P. 833 [81].

In [82] and [83] signal-based approaches for estimating the echo parameters have been proposed. Similarly, Nuenes et al. [84] proposed a signal-based procedure for estimating the echo delay and echo gain. The proposed procedure was then used to develop a parametric objective quality metric for evaluating the quality of full band speech degraded by acoustic echo; the proposed model is a result of mapping the delay and echo gain into the mean subjective score. A high correlation of 94 % between the predicted and actual subjective scores was reported.

While parametric models can easily be placed in the network elements and terminals, they do not have the efficiency of the signal-based models in predicting the perceived speech quality [35]. Among the signal-based objective models, the ones which are based on modelling the human auditory system are more attractive [36]. Several auditory-based speech quality metrics have been proposed [54, 53, 85, 46] and their performances have been evaluated in different types of impairments caused by, for example, noise reduction algorithms [43, 86, 44], coding process [46] etc. It is worthwhile to note here that the ITU standard PESQ [37] is the de facto standard for speech quality assessment with its wideband version standardized in ITU 862.2 [38].

Only a limited number of these models have been used for studying the effect of echo. For example, Biscainho et al. [87] investigated the impact of echo degradations on the perceptual quality of speech sampled at 48 kHz through objective and subjective speech quality measurements. Subjective scores have been collected via listening tests according to ITU-T P. 800. The objective model proposed by the authors used one of the metrics from ITU-R BS.1387 [85] for its feature extraction part. A high correlation coefficient between objective and subjective scores was reported.

In another study, using a narrowband speech database that contained degradations caused by noise, echo, delay, packet loss in a conversational context, Guéguin et al. [35] developed an objective metric for evaluating the quality of speech. The proposed metric, was a combination of PESQ (an objective model for listening quality test), PESQM (an objective model for talking quality test) [41] and delay in the given condition. The performance of both PESQ and PESQM was found very good in terms of correlation coefficient and mean absolute error.

The goal of the research presented in this Chapter was to investigate the effect of impairments caused by echo and delay on the quality of wideband speech in the listening context, subjectively and objectively. In particular, subjective listening test was conducted to evaluate the effect of delay, echo and different echo path models on the speech quality. Further-

Table 4.1: List of the sentences used for the echo quality database.

Filename	Speaker	Gender	Sentence text
S01.wav	MA	M	Glue the sheet to the dark blue background. These days a chicken leg is a rare dish. Rice is often served in round bowls.
S02.wav	MC	M	The juice of lemons makes fine punch. The box was thrown beside the parked truck. Four hours of steady work faced us.
S03.wav	FD	F	A young child should not suffer fright. Add the column and put the sum here. She has a smart way of wearing clothes.
S04.wav	FD	F	Where were they when the noise started. Bring your best compass to the third class. They could laugh although they were sad.

more, the performance of the LPD+BMLS objective model (expounded in Chapter 2, and applied in Chapter 3) in predicting the quality of speech degraded by echo is explored.

The remainder of this Chapter is organized as follows: Section 4.2 describes the creation of customized database and the subjective data collection for evaluating the effect of echo and delay on the wideband speech. Section 4.3 reports the analyses of subjective data, which is followed by the objective evaluation, presented in Section 4.4. The Chapter is summarized in Section 4.5.

4.2 Echo Quality Database

A custom database was created for this project based on 4 clean speech samples produced by one female and two male speakers. Each clean speech sample is a concatenation of three sentences, taken from the TSP speech database [72]. As the average length of speech samples in TSP speech database was 2.372 s, this concatenation was necessary to ensure that the speech samples were long enough for representing impairments caused by varied echo and delay. The average length of the samples in the database was 6.504 s.

The original sampling rate of TSP speech database was 48 kHz which was down-sampled to 16 kHz for this study. The list of sentences is given in Table 4.1. Each speech sample in the database was generated by adding the echo signal to the clean signal; the echo signal was produced by adding delay to the clean signal and then passing it through echo path model. Four echo path models and 19 test conditions including different delay and ERL (Echo Return Loss) values were used in generating the echo signals.

Table 4.2: Echo path models used for the echo quality database

Label	Description
Handset1	Room impulse response (office), Handset mode, Length:50 ms
Handset2	RIM, Acoustic echo path, Handset mode, Length:16 ms
Handsfree1	Room impulse response (office), Hands-free mode, Length:50 ms
Handsfree2	Room impulse response (meeting room), Hands-free mode, Length:50 ms

Of the four echo path models, one was obtained from Research In Motion (RIM) which is the acoustic echo path in handset mode and the other three were taken from Aachen Impulse Response database, known as "AIR database" [88, 89]. AIR database is a set of room impulse responses (RIRs) measured in several indoor environments such as office, meeting room, corridor etc. Measurements were performed for two positions of the phone "Hand-Held Position (HHP)", and "Hands-free Reference Point (HFRP)" and impulse responses between an artificial mouth and microphone were measured and reported as RIRs. The original measurements were done at 48 kHz which were down-sampled to 16 kHz for the present study. The original length of RIRs was 3 sec, only the first 50 ms of each RIR was used in this study. The magnitude response of the echo-path models was normalized such that it did not exceed 0 dB. A brief description about all the models used in this work is given in Table 4.2. Figure 4.2 to Figure 4.5 show the impulse and frequency responses of these models.

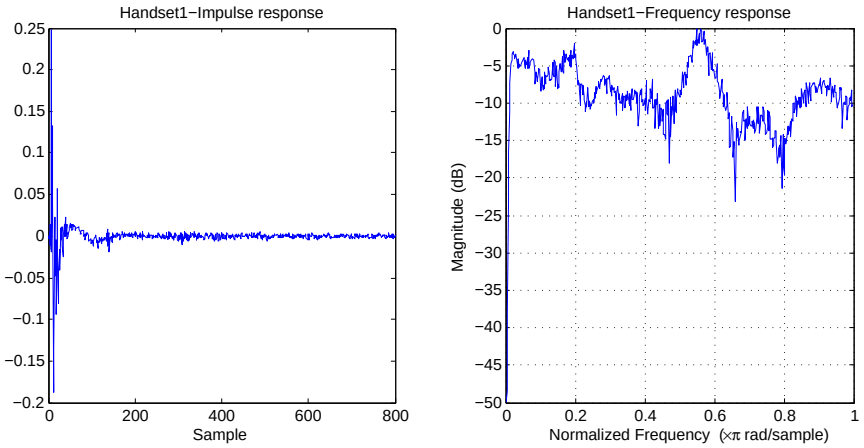

Figure 4.2: Impulse response and magnitude response for Handset1

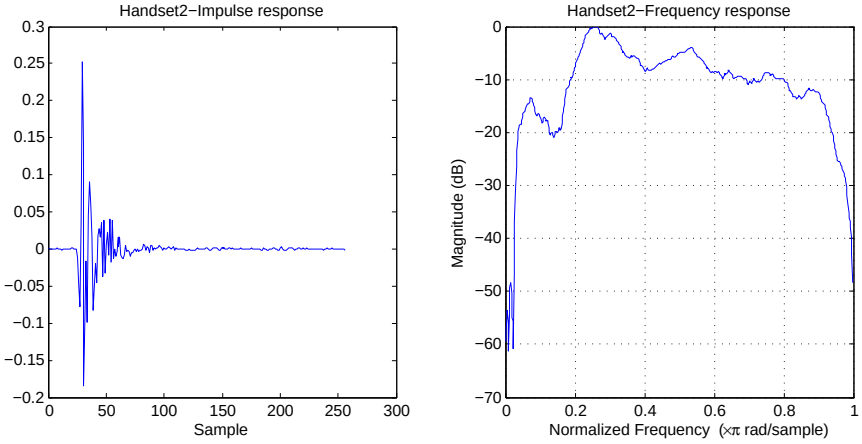

Figure 4.3: Impulse response and magnitude response for Handset2

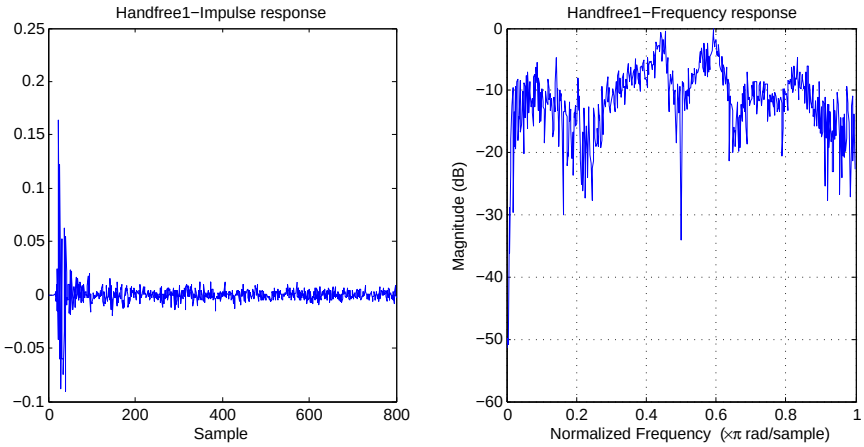

Figure 4.4: Impulse response and magnitude response for Handsfree1

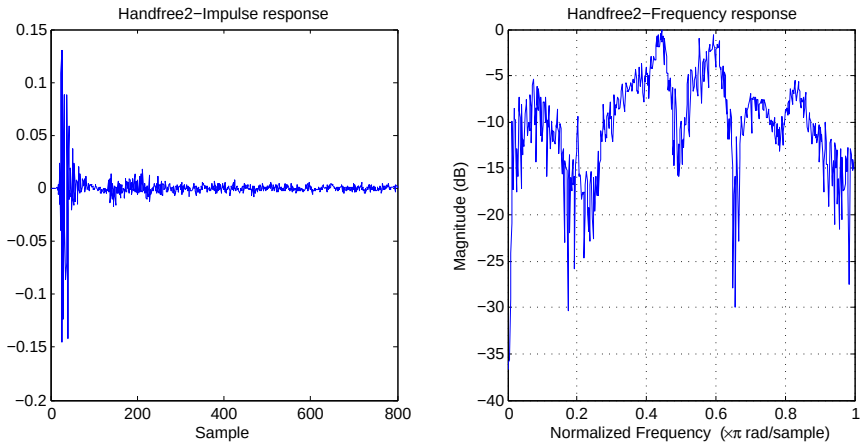

Figure 4.5: Impulse response and magnitude response for Handsfree2

Table 4.3: Test conditions

ERL(dB)	Delay (ms)	Description
60	400	Acceptable audible echo
40	50	Acceptable audible echo
46	200	on the border("Acceptable audible echo" and "Unacceptable audible echo")
20	0	Unacceptable audible echo
46	300	Unacceptable audible echo
20	150	Unacceptable audible echo
30	150	Unacceptable audible echo
35	200	Unacceptable audible echo
20	200	Unacceptable audible echo
10	200	Unacceptable audible echo
30	300	Unacceptable audible echo
20	300	Unacceptable audible echo
10	300	Unacceptable audible echo
0	300	Unacceptable audible echo
25	400	Unacceptable audible echo
10	400	Unacceptable audible echo
5	500	Unacceptable audible echo
10	600	Unacceptable audible echo
0	600	Unacceptable audible echo

Nineteen test conditions, shown in Table 4.3, were selected based on a report presented by Nortel [79] on user's perception of talker echo, and the descriptions given for each condition in the table show those users's rating for that condition. The test conditions were chosen such that they cover the whole range of echo quality including "unacceptable audible echo" and "acceptable audible echo".

Subjective data collection was performed in two steps for reasons described below. In the first study, the first nine test conditions were used in developing the database and subjective quality ratings were obtained (complete details of subjective data collection are given later in this Section). Subsequent analysis of the data revealed that the entire range of quality was not covered. Figure 4.6 depicts the scatter plot of the condition-averaged ratings across the test conditions in the first study. In this Figure, the x-axis shows the number assigned to each test condition. This assignment, shown in Table 4.4 was made such that the test condition which got the lowest rating had the lowest index and so on. The goal of depicting the scatter plot in this way was to display how different test conditions were rated and whether the whole range of quality had been covered consistently.

It can be noticed from Figure 4.6 that of the nine test conditions, five of them were rated above 90, and the other four were rated between 30 and 90. In fact, even some of the test-conditions which were labeled as "unacceptable audible echo" in Table 4.3, were rated

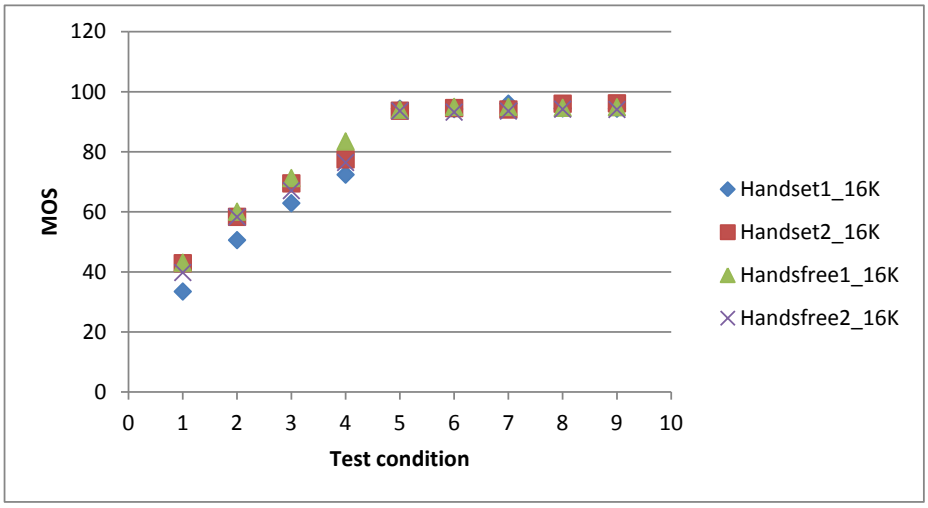

Figure 4.6: Plot of the condition-averaged ratings across test conditions of the first study for 16 kHz

Table 4.4: Test conditions of the first study – as labeled in x-axis of Figure 4.6

Index	1	2	3	4	5	6	7	8	9
Delay-ERL	200ms-10dB	300ms-20dB	400ms-25dB	150ms-30dB	200ms-46dB	300ms-46dB	400ms-60dB	0ms-20dB	50ms-40dB

above 90 and there were not enough ratings between 20 and 90.

This result was not surprising, because as mentioned earlier, the labels in Table 4.3 were based on a talker echo loudness rating study; i.e., a specific amount of echo in talking-and-listening test is more perceivable as compared to listening test alone, and that is why the echo which was labeled as “unacceptable audible echo” was rated almost the same as the “acceptable audible echo” ones in our listening test.

To develop a model for predicting the quality scores, the model needs to be trained by subjective quality scores which cover the whole range of quality. For this purpose, it was decided to perform the second study using additional new ERL and delay values, and the last ten test conditions shown in Table 4.3 were chosen for this purpose. All of these test conditions were chosen such that they produce “Unacceptable audible echo” in talking-and-listening test.

Figure 4.7 shows the scatter plot of the condition-averaged ratings across test conditions of both studies. It can be seen that the whole range of quality was now being covered by the test conditions chosen for both studies. Table 4.5 shows the test conditions related to the indices on the x-axis of Figure 4.7.

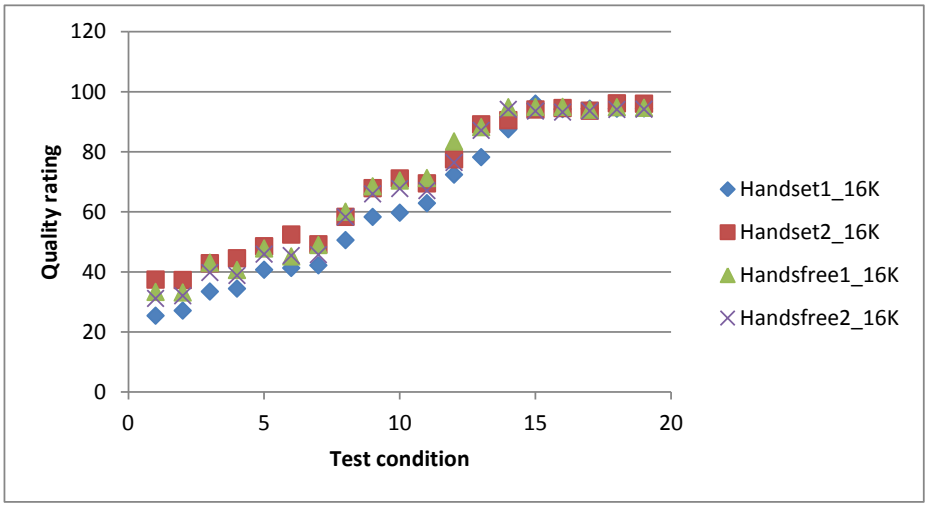

Figure 4.7: Plot of the condition-averaged ratings across test conditions of both studies for 16 kHz sampling rate

Table 4.5: Test conditions for both studies - as labeled in x-axis Figure 4.7

Index	Test condition (delay-ERL)	Index	Test condition (delay-ERL)	Index	Test condition (delay-ERL)
1	300ms-0dB	8	300ms-20dB	15	400ms-60dB
2	600ms-0dB	9	200ms-20dB	16	300ms-46dB
3	200ms-10dB	10	150ms-20dB	17	200ms-46dB
4	500ms-5dB	11	400ms-25dB	18	50ms-40dB
5	300ms-10dB	12	150ms-30dB	19	0ms-20dB
6	400ms-10dB	13	300ms-30dB		
7	600ms-10dB	14	200ms-35dB		

With this background, additional details on subjective data collection and data analysis for both studies are given below.

4.2.1 Subjective Data Collection

A total of 17 subjects participated in the two experiments described earlier, ten in each one; and three subjects participated in both studies.

In the first study, all 10 participants were students at Western, 3 males and 7 females; and in the second one, out of 2 male and 8 female participants, 6 were audiology students and

the other 4 ones were in speech language pathology program. None of the participants had explicit knowledge of the stimuli beforehand, but as mentioned earlier, three subjects participated in both studies. English was the first language for all participants. Hearing thresholds were measured using TDH headphones at audiometric frequencies between 250-8000 Hz. All fell within the normal range (below 25 dB HL) at each test frequency except one participant, who had a mild loss (35 dB HL threshold) at 1000 Hz only. At all other frequencies, this person's hearing was normal, and overall would still be considered to have normal hearing.

For the sound quality ratings task, participants were seated in a double-walled sound booth and listened to stimuli over Sennheiser HDA 200 headphones. Subjects impression of the echo impairment in the test samples was gathered in a MUSHRA test (see a screenshot of the software in Figure 4.8) [34] and the sound level was adjusted to their comfortable listening level. There were 16 different screens[1] that participants worked through and each screen had 10 and 11 samples for the first and second studies respectively. One of the icons, which is on the most left side of the screen is the reference signal, and the other 9 (in the first study) and 10 (in the second study) speaker icons were randomly associated with speech signal degraded by echo at the nine condition tests (the first study) and the ten conditions (the second study) given in Table 4.3. Participants were told that they would hear sentences that have been degraded by echo and they could listen to each sentence as many times as they liked by clicking on the speaker icons. Each sentence was then displayed at the bottom of the screen so that participants could see what the sentence was.

Participants were instructed to rate each stimulus using the sliders, by paying particular attention to the impairments due to delay and echo. They were also asked to rate the stimulus based on its comparison to the reference signal and the other stimuli on the screen. In this way, not only each signal was rated in comparison with the reference speech, but also the differences between the stimuli at different conditions were taken into account. When participants were satisfied with the rating for each speaker, they were instructed to click "Save and Continue" button to get to the next screen.

Subjective and objective analyses were performed to investigate and compare the effect of different echo path models and test conditions. Since two different groups of subjects rated the quality of each database, separate statistical analysis was conducted on the quality ratings of each study.

[1]There is one screen per sentence per each echo path model. Since there are four echo path models, the total number of screens would be 4×4 Sentences=16.

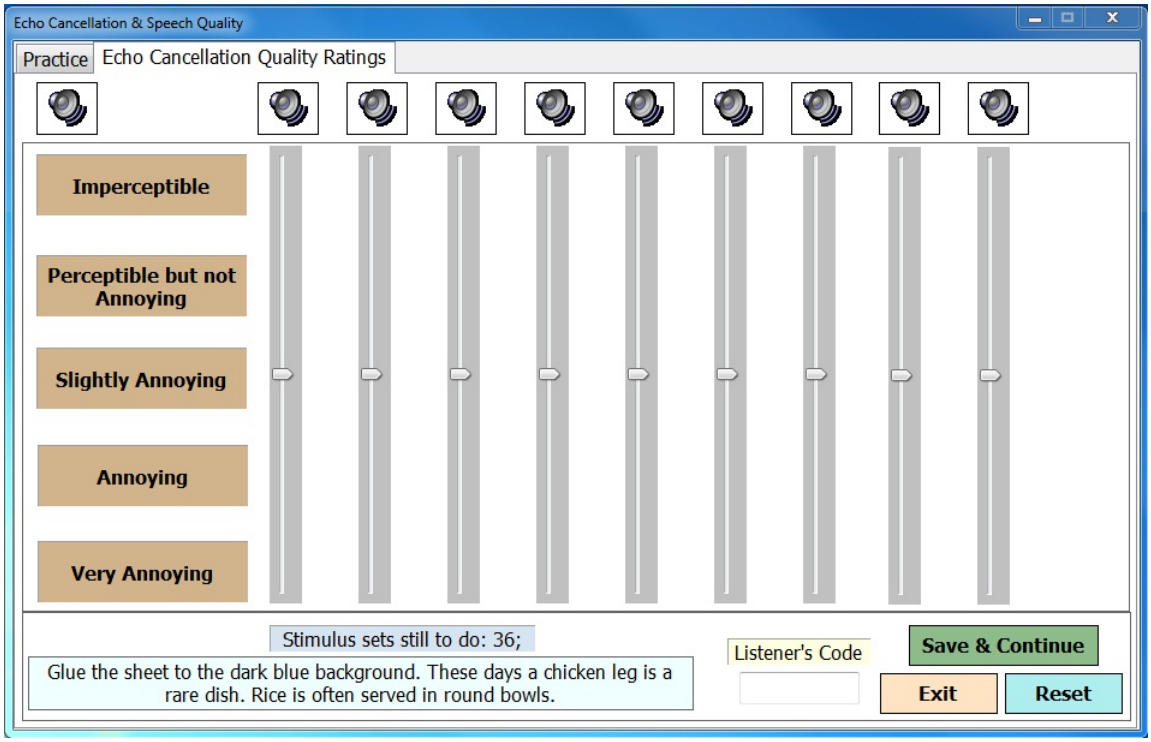

Figure 4.8: Screenshot of the MUSHRA quality ratings software

4.3 Subjective Analysis

Reliability of the ratings

The quality ratings were first analyzed for the inter-rater reliability. The consistency of ratings among the 10 participants (in each study) was measured using Cronbach's α and Intraclass Correlation Coefficient (ICC) using SPSS statistical software package, Version 19.0. The Cronbach's α and the range of average measure ICC are 0.983 and 0.967 – 0.980 for the first study and 0.981 and 0.956 – 0.972 for the second one. These numbers show a high degree of reliability and very good agreement among listeners on the quality scores.

Averaged ratings (plots)

Figure 4.9 depicts the average speech quality ratings of different echo path models at 19 test conditions. The error bars in these figures represent one standard deviation. In order to quantify differences among all conditions, a thorough statistical analysis was performed.

A split-plot repeated measures ANOVA was first performed in SPSS software with test conditions and echo path models as the "within subject factors". Significant two-way interactions were found between test conditions of the first study and echo path models

$(F(4.796, 43.167) = 4.497\ p < 0.05)$.

Using the FDR (false discovery rate) control method [73], multiple comparisons were performed to assess the significance of the differences between quality scores obtained from different test conditions of the first study for each echo path model. Tables 4.6-4.9 show the result of this test for all four echo path models. These tables present data that will help determine how different test condition have been categorised for different echo path models.

As can be seen from the tables, the results of the analysis demonstrated six distinct groupings of the nine test conditions for "Handset2", "Handsfree1" and "Handsfree2". This grouping is a little different for "Handset1" model. There were 7 subsets for this model. The difference between "400ms-60dB" and "300ms-46dB" conditions for this model was the source of the difference between this model and the other ones, since the other subsets were same for all the echo path models.

These results showed that the shape of the frequency response of the echo path models does not have significant effect on the speech quality, and delay and ERL parameters are the ones which have the most effect on the speech quality.

Looking at the test conditions in the subsets, it is possible to compare the effect of ERL and delay parameters on the speech quality. The results showed that the parameter "ERL" had more effect on the quality than the amount of delay. The test conditions with the large ERL values, and different delay values, have been classified as the subset with the highest quality ratings. For example, comparing the conditions "200ms-10dB" and "200ms-46dB" shows that the same amount of delay, but different ERL values resulted in the echo-corrupted speech samples with the worst and best subjective quality scores. It is almost same for the conditions "300ms-20dB" and "300ms-46dB' and also "400ms-25dB" and "400ms-60dB".

The other way this data could be analyzed was to look at each test condition individually and perform FDR control method to see if there were any differences between the echo path models for that condition and how they were categorized for that test condition. However, because of the way that the echo path models were normalized, the result of this analysis may not be valid. This is discussed in further detail in the following section.

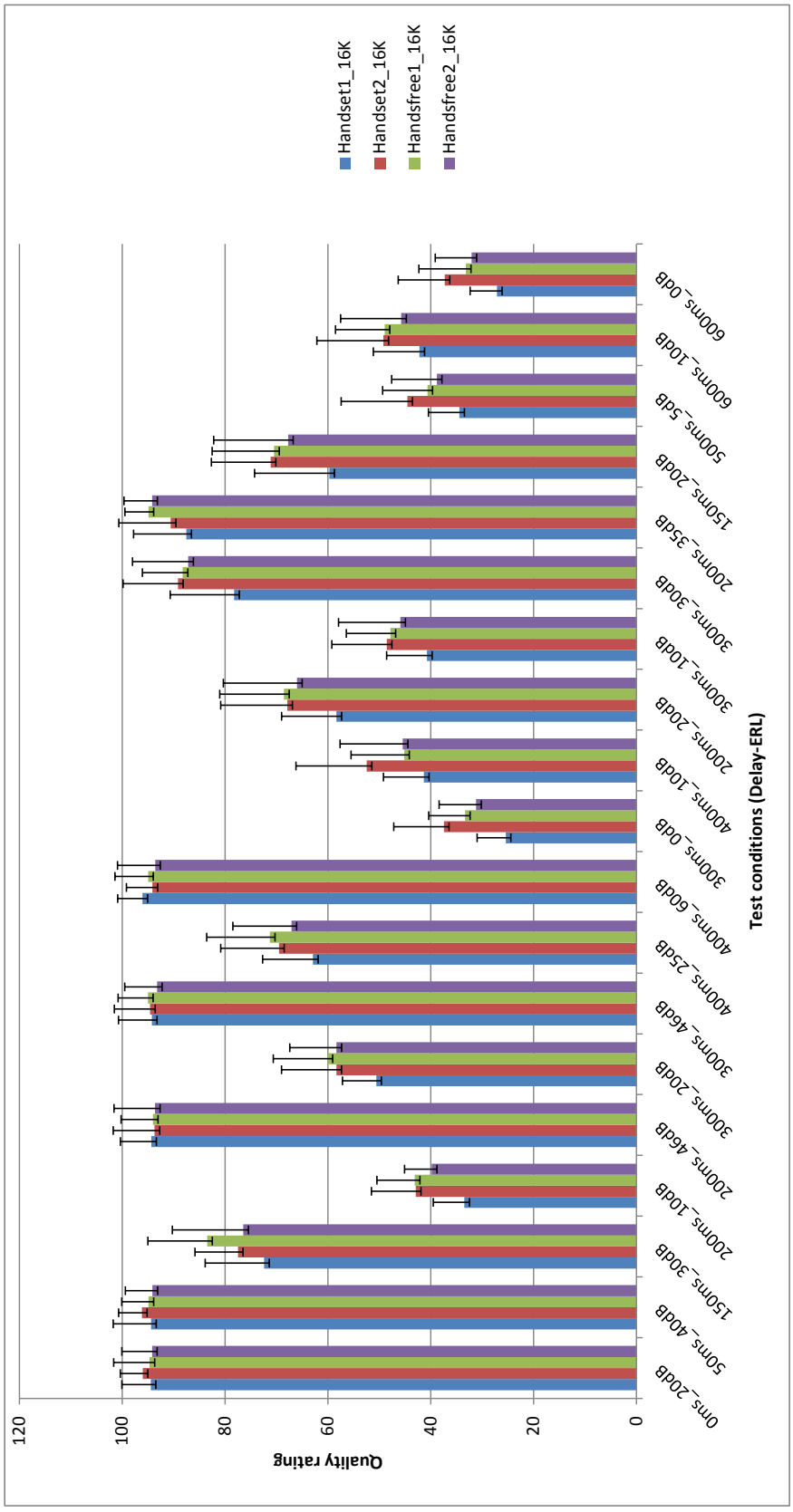

Figure 4.9: Speech quality ratings for different echo path models at several delay and ERL values – Sampling rate: 16 kHz

Table 4.6: Results of the post-hoc test (Handset1)

Test conditions	Subset for $\alpha = 0.05$					
	1	2	3	4	5	6
'200ms-10dB'	33.475					
'300ms-20dB'		50.6				
'400ms-25dB'			62.9			
'150ms-30dB'				72.4		
'300ms-46dB'					94.225	
'200ms-46dB'					94.35	94.35
'50ms-40dB'					94.375	94.375
'0ms-20dB'					94.45	94.45
'400ms-60dB'						96.05

Table 4.7: Results of the post-hoc test (Handset2)

Test conditions	Subset for $\alpha = 0.05$				
	1	2	3	4	5
'200ms-10dB'	43.1				
'300ms-20dB'		60.075			
'400ms-25dB'			71.275		
'150ms-30dB'				83.475	
'200ms-46dB'					94.025
'400ms-60dB'					94.675
'300ms-46dB'					94.875
'0ms-20dB'					94.95
'50ms-40dB'					94.975

Table 4.8: Results of the post-hoc test (Handsfree1)

Test conditions	Subset for $\alpha = 0.05$				
	1	2	3	4	5
'200ms-10dB'	43.1				
'300ms-20dB'		60.075			
'400ms-25dB'			71.275		
'150ms-30dB'				83.475	
'200ms-46dB'					94.025
'0ms-20dB'					94.675
'50ms-40dB'					94.875
'400ms-60dB'					94.95
'300ms-46dB'					94.975

Table 4.9: Results of the post-hoc test (Handsfree2)

Test conditions	Subset for $\alpha = 0.05$				
	1	2	3	4	5
'200ms-10dB'	39.8				
'300ms-20dB'		58.325			
'400ms-25dB'			67.075		
'150ms-30dB'				76.45	
'300ms-46dB'					93.225
'400ms-60dB'					93.575
'200ms-46dB'					93.625
'50ms-40dB'					94.1
'0ms-20dB'					94.2

Real ERL values

As mentioned earlier, the echo path models were normalized such that the maximum value of the echo path frequency response was 0 dB, and after applying the ERL value, that maximum point was limited to the chosen ERL value.

The scaling factor used for normalizing the echo path model depends on the input signal used for the test. ITU-T G.168 [90] uses different scaling factors for different input signals such as Composite Source Signal (CSS), white noise and tone signal. The normalization method used for the first two studies is typically used when the input is a tonal signal, and as such was not a proper normalization method for speech. Therefore, real ERL values resulting from speech input signals were calculated. The active level of the reference and the echo signal was measured according to the ITU-T P.56 [91] and the difference between these two levels represented the real ERL value.

Table 4.10 shows the real ERL values and also the ERL values reported as the test conditions when the database was created. Real ERL values were the result of averaging the differences between the active speech and echo levels across the four speech samples used for making the database. Comparison between the real ERL values and the chosen ERL values shows that the real attenuation applied by the filters is more than the ERL values specified in the test conditions; in addition, the amount of attenuation is not same across the echo path models.

Table 4.10 shows that the effective attenuation applied by "Handset1" model is less than the level applied by the other models. So, at a specified ERL value, the echo produced by this model was more perceivable in compared to the other ones.

Table 4.10: Comparison of the ERL values of the speech input signal and the ERL values used for the input tone signal - sampling rate =16 kHz

ERL(dB) (tone signal)	0	5	10	20	25	30	40	46	60
Echo path models	(Active level of the clean signal-Active level of the echo signal) (dB)								
Handset1	5.87	10.87	15.87	25.87	30.87	35.87	45.87	51.87	65.87
Handset2	8.41	13.41	18.41	28.41	33.41	38.41	48.41	54.41	68.41
Handsfree1	9.65	14.65	19.65	29.65	34.65	39.65	49.65	55.65	69.65
Handsfree2	8.00	13.00	18.00	28.00	33.00	38.00	48.00	54.00	68.00

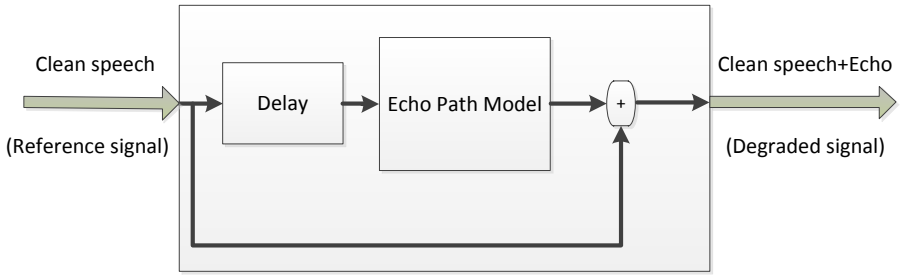

Figure 4.10: The reference and degraded signals used by the objective models

4.4 Objective Analysis

Using the echo quality subjective scores garnered from the listening test, the performance of WPESQ and LPD-BMLS objective estimation models was investigated. As described earlier, both WPESQ and LPD+BMLS are intrusive models and require both the reference and the degraded signals for predicting the quality score. In this work the clean signal was used as the reference signal and the speech signal plus the echo was used as the degraded signal. These two signals are shown in Figure 4.10.

Correlation coefficients and standard errors of estimation were used for evaluating the performance of these two models. Two types of correlation analysis were employed. In the first analysis, objective scores were computed for each speech sample which was then used for correlation analysis; i.e. all 304 scores (4×4(echo path model)$\times$ 19(test conditions)) were used for evaluating each objective measure.

In the second analysis, the average score at each condition was used for correlation analysis. In fact, after calculating the objective scores for all speech samples, the scores of all 4 speech samples at a specific condition (same echo path model and test condition) were averaged. These condition-averaged scores were later used for correlation analysis. As there were 76 (4(echo path model)$\times$ 19(test conditions)) conditions, 76 pairs of subjective

Table 4.11: Correlation coefficient and standard error of estimation for WPESQ and LPD-BMLS

Objective measure	ρ	σ_e	ρ_ave	σ_e_ave
WPESQ	0.9704	5.7064	0.9809	4.5837
LPD+BMLS	0.9802	4.6846	0.9849	4.0788

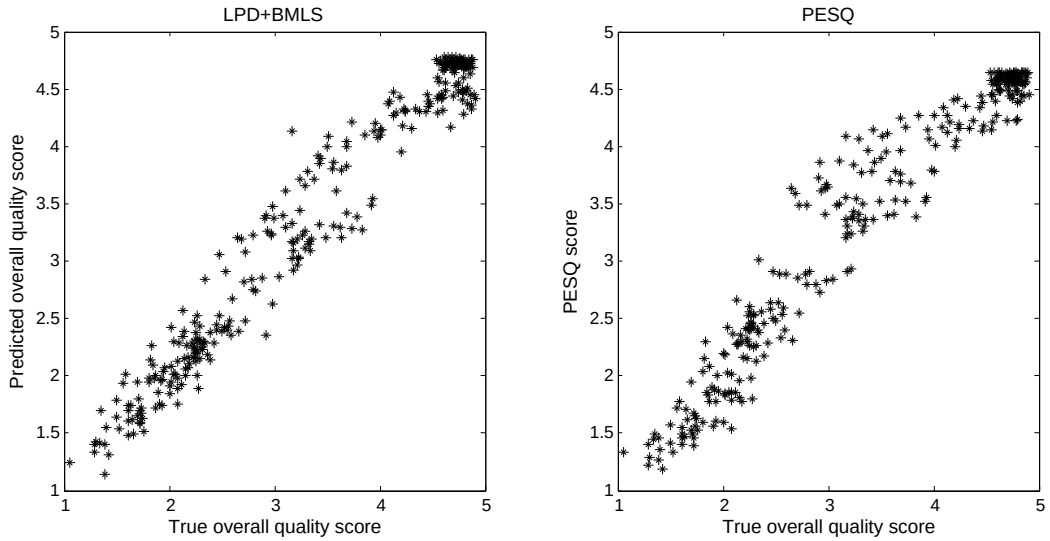

Figure 4.11: Scatter plot of the predicted and actual overall quality scores for LPD-BMLS and WPESQ and

and objective scores were available for condition-average analysis.

Table 4.11 shows the correlation coefficients and the standard errors of estimation for WPESQ and LPD-BMLS. Both models exhibited high correlation with the subjective scores. Figure 4.11 depicts the scatter plot of the predicted values of the overall quality scores versus the actual scores for WPESQ and LPD-BMLS.

Steiger's Z test [74] was performed to test if the difference between LPD+BML and WPESQ was significant. The result of the test showed that there was significant difference between WPESQ and LPD+BMLS for per-sample analysis ($Z = 5.27, p < 0.05$), but their performances were statistically similar for condition-average analysis ($Z = 1.34, p > 0.05$).

4.4.1 Application of LPD-BMLS

Since LPD+BMLS correlated highly with the subjective scores, it can be used to evaluate the performance of echo cancellers (ECs) and echo suppressors (ESs). In the present, the LPD+BMLS metric was applied to state-of-the-art EC and ES algorithms, as detailed

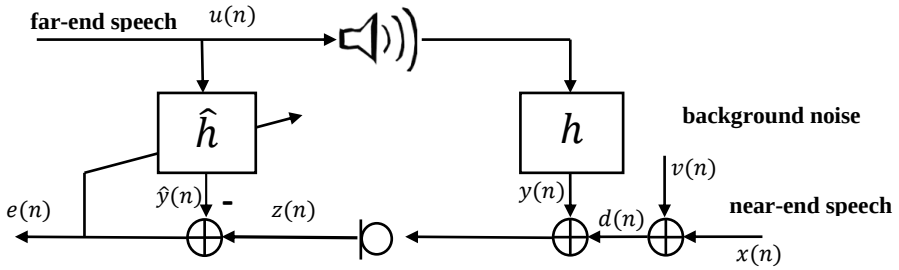

Figure 4.12: Block diagram of a standard echo canceller.

below.

4.4.2 Echo Canceller (EC)

Figure 4.12 depicts the block diagram of the echo canceller in telephony applications. Here, the far-end speech is played out through the speaker, and the microphone receives a mixture of the near-end speech, background noise, and a version of the far-end speech. The block "h" represents the time-varying echo path between the speaker and the microphone.

The time-varying echo path is typically estimated using an adaptive filter, $\hat{h}$. This is commonly achieved through the normalized Least Mean Square (NLMS) adaptive algorithm, where the filter weights are iteratively estimated using the following equation:

$$\hat{h}(n+1) = \hat{h}(n) + \frac{\mu}{\delta + ||\mathbf{u}(n)||^2}\mathbf{u}(n)e^{\star}(n) \tag{4.1}$$

where is the convergence constant and δ the regularization parameter. The NLMS-based echo cancellation algorithms can be implemented in either time-domain or frequency-domain [92]. Moreover, to speed up the convergence of the adaptive filter and to reduce the computational complexity associated with long filter lengths used to model the echo impulse response, various flavours of the basic NLMS algorithm including subband adaptive algorithms, affine projection algorithms, variable step size NLMS algorithms, and Frequency-domain adaptive algorithms such as fastLMS (FLMS) algorithm have been proposed [92, 93, 94].

The echo canceller evaluated is this work is based on flexible multidelay block frequency domain (MDF) adaptive filter [95]. This filter has some advantages over the standard

NLMS based algorithms mainly in terms of lower memory and hardware usage and being computationally more efficient.

4.4.3 Echo Suppressor (ES)

The echo canceller is usually followed by another block called nonlinear processor (NLP) or echo suppressor (ES). The echo signal is first cancelled by the echo canceller; the echo that remains after this operation, called the residual echo, is removed by the suppressor. An example of an ES is an analog centre clipper; all signal levels below a defined threshold are suppressed and forced to some minimum value and higher level content is allowed to pass unaffected [78].

The echo suppressor used in this work is based on a simple switched loss algorithm which controls the level of the signal at each path and inserts loss in the path with the signal level below its specified threshold.

4.4.4 EC/ES Evaluation

LPD+BMLS score was calculated for speech samples before and after applying the echo suppressor and echo canceller. There are four different states: (a) neither ES nor EC was active, (b) Only ES was active, (c) only EC was active, and (d) both were active.

Figure 4.13 shows the scatter plot of LPD+BMLS score after each of these states versus the LPD+BMLS score before using echo canceller and echo suppressor. It is evident from this figure, the quality scores did improve after using ES (Figure 4.13-b) compared to Figure 4.13-a, when there is no ES and EC is applied. LPD+BMLS scores were even higher after deactivating ES and activating EC, Figure 4.13-c. These Figures also reveal that EC alone is more powerful than ES alone for cancelling the echo. Finally, using both ES and EC results in an almost equal LPD+BMLS score across all test conditions. This means that a combination of EC and ES will mitigate echo at even annoying test conditions and increase the quality scores of the speech samples at those condition as if no echo exists in the samples.

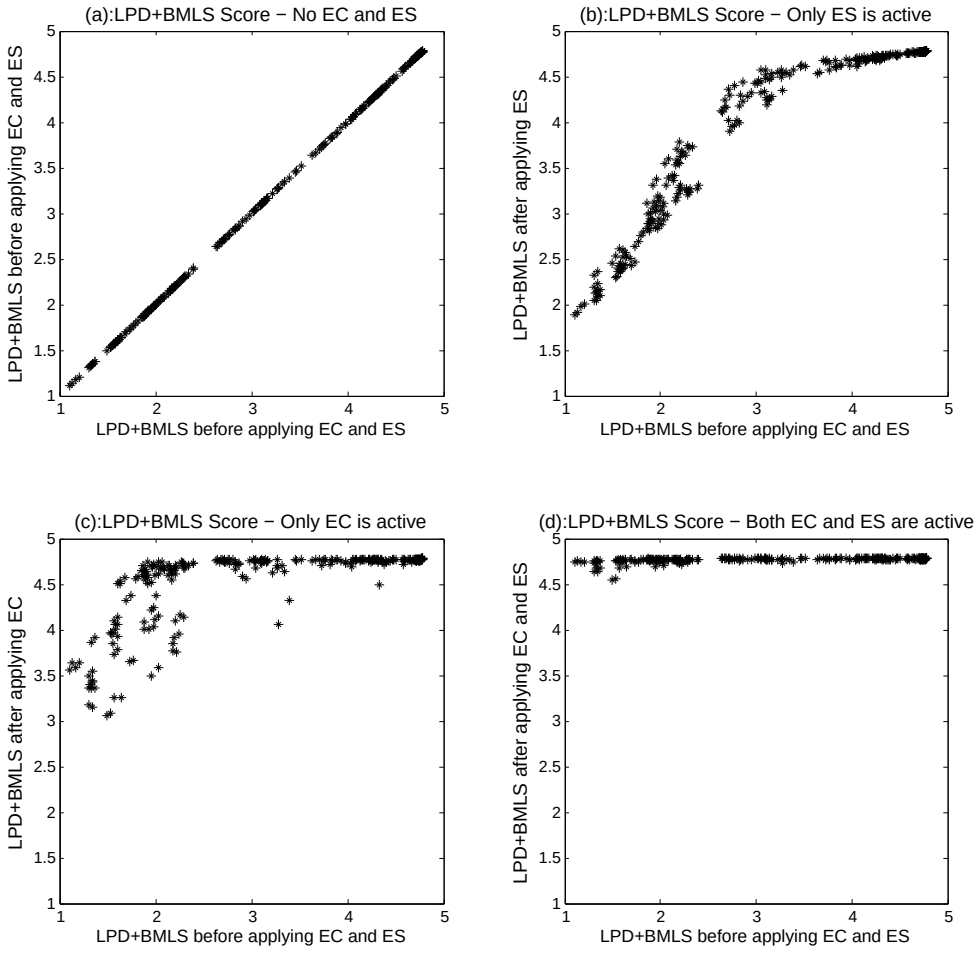

Figure 4.13: Scatter plot of the LPD+BMLS score computed (a) before using ES and EC, (b) after applying ES, (c) after applying EC, (d) after using both EC and ES (sampling rate = 16 kHz)

4.5 Summary

In this Chapter, the effect of delay and echo on the quality of speech in the listening context, was evaluated subjectively and objectively. A listening test was performed and MUSHRA software was used for subjective data collection. The effect of four echo path models were evaluated. Several end-to-end delay and ERL values were used to generate echo at different levels of annoyance including not noticeable echo to annoying echo.

WPESQ and LPD-BMLS were used for predicting the quality of the speech signals corrupted by the echo. There was a high correlation between the subjective scores and objective scores predicted by these two models. The LPD-BMLS model was also used to evaluate the effect of an echo canceller and echo suppressor. Results from this analysis

showed that neither algorithm was able to tackle echo at all conditions, and a combination of EC and ES was required to substantially reduce echo.

Chapter 5

Echo Quality in Conversation Context

5.1 Introduction

As is evident from previous chapters, subjective evaluation of speech quality is integral part of benchmarking signal processing algorithms for telecommunication applications. A variety of subjective test procedures have been recommended by ITU-T for evaluating specific signal processing algorithms. For example, the ITU-T P.835 standard specifically describes the subjective testing procedure for evaluating the quality of speech degraded by noise reduction algorithms [30]. The rating procedure in this specification was designed to reduce the listeners' uncertainty and confusion to score the quality of enhanced speech by using separate rating scales for evaluating "speech", "background noise" and "speech + background noise".

Four subjective testing methods have also been suggested in ITU-T recommendation P.831 [31] for performance evaluation of echo canceller algorithms. These tests include: conversational test, talking-and-listening test and third-party listening test-type A and B. It also has been recommended that talking-and-listening tests, and listening only tests should not be done in isolation and should be followed by conversational test which involves interactions between subjects [31].

Each method of subjective test has advantages and disadvantages that make it suitable for certain applications. For example, while conversation tests represent the most realistic way of speech quality assessment, listening test may be the only possible assessment option when a new transmission system is being developed. Besides that, not all of the subjective tests are applicable for all forms of degradations affecting speech quality. Some of the degrading factors include: packet loss, echo, delay, low bit rate coding and etc.

One of the main limitations of talking-and-listening and listening only tests is that they cannot be used for studying the impact of conversational related parameters and effects such as delay and double talk, as the impairments caused by these parameters can only be experienced when there is interaction between the two subjects [31, 96]. The most realistic way to study the effect of the parameters like delay, is to perform a conversational test.

In several studies, the effects of delay and echo on the quality of the transmission system was evaluated subjectively in the conversational context [97, 98, 99, 100]. In [97], speech quality degradation caused by pure delay, delay plus echo and echo suppressors was investigated. For this purpose, an experimental circuit was inserted into naturally occurring telephone conversations and subjects were told that some of their calls would be routed through a simulated satellite circuit. The users did not know which calls were affected and also what changes were made in their circuits. They were also instructed to reject a call and restore the standard circuit if they found any call unsatisfactory for the normal use. The number of calls rejected by the users because of having unsatisfactory quality for the normal use is an indicator of transmission quality. In this study the effect of exposure to the delay, adding noise and loss to the circuit and using different echo suppressors were investigated. The rejection rate was increased after exposing to the delay circuits and also with increase in return loss and noise level. Among the reasons of rejecting a call, echo got the first rank followed by noise, low volume, chopping and delay respectively.

Another study by the same people [98] showed that users are rarely disturbed by 600 ms and 1200 ms pure delays and even exposing to longer delay value (2400 ms) does not make any increase in the rate of call rejection.

Similarly in [99], the effect of delay in trans-Atlantic calls was studied. A number of different echo suppressors were also tested. Customers were interviewed after making calls over the simulated circuit. Results show that the quality of communication systems with echo suppressors decreases with increasing transmission delay and also while none of the echo suppressors showed superior performance for all delay values; but some of them appear better for longer delays.

In other work, the effect of unsuppressed echo in long-delay telephone conversation was subjectively evaluated [100]; 40 test conditions including four delay, five values for echo return loss and two echo-path frequency characteristics (flat and rising with increasing frequency) were used for this purpose. A simulated long-delay telephone link was used for subjective data collection. Each pair of subjects participated in ten test conditions; they were asked to converse over the simulated telephone link and then rate the conversation

on the MOS scale and also mention if they had any difficulty during the call. Results showed that:1) for both echo-path frequency characteristics, the MOS score decreased with increasing delay, 2) for high value of echo return loss (50 dB), the MOS decreases slightly with increasing delay.

While subjective methods are the most reliable way for evaluating the speech quality, they are also expensive and time-consuming; therefore having an objective model which can estimate the speech quality using the information of the system (parametric model) or the speech signal (signal-based models) is attractive. The state of the art parametric model is E-model, standardized by ITU in recommendation G. 107 [18] as a transmission planning tool. While parametric models can easily be placed in the network elements and terminals, they do not have the efficiency of the signal-based models in predicting the perceived speech quality [35].

Most of the proposed signal-based models are listening-quality models which are used for evaluating speech quality in the listening context. There are limited number of models for the conversational contexts.

An objective talking-quality model, called the perceptual echo and sidetone quality measure (PESQM), has been developed in [41]; different types of degradation including single echo, sidetone distortion and background noise were used for developing the model. In talking quality experiments, subjects evaluate the quality of their own voice while they are speaking. In this study all subjective scores were collected when subjects were actively speaking. Six talking quality tests were carried out based on ITU-T P.800 [27].

PESQM along with PESQ (an objective model for listening quality test) [37] was used by Guéguin et al. [35] for developing an objective model for predicting the speech quality in conversation context. Using a narrowband speech database that contained degradations caused by noise, echo, delay, packet loss in a conversational context, Guéguin et al. [35] collected speech quality ratings from a group of listeners. The subjective test was a conversation test involving two non-expert subjects (subjects A & B) and consisted of three parts: conversational, talking and listening. At the end of each part, both subjects rated the quality of the communication according to an absolute category rating (ACR) opinion scale. The objective speech quality measure was computed as a combination of PESQ, PESQM and a known delay in a given condition. High correlations were reported between objective and subjective scores across noisy and echo conditions.

In summary, conversational speech quality assessment is an important subtopic within broader topic of speech quality evaluation of telecommunication devices, networks, and al-

gorithms. Conversational speech quality assessment can probe the performance of telecommunication systems along dimensions not assessable through listening-only or talking-and-listening tests. The paucity of data and results for wideband conversational speech quality assessment, both objective and subjective, motivated the research presented in this Chapter.

To perform the conversation test, it is imperative to use a realtime telephone line simulator. As such, the first goal of this research project was to develop a custom software module to simulate a real world telephone conversation. This realtime simulator was then employed for systematically investigating the effect of delay and echo in the conversational context, both subjectively and objectively.

The rest of the Chapter is organized as follows; the development of the realtime simulator and its evaluation are presented in Section 5.2. Section 5.3 details the procedures followed for performing the subjective test including calibration and data collection. The objective evaluation procedure and results are presented in Section 5.4, and finally the chapter is summarized in Section 5.5.

5.2 Realtime Simulator

Figure 5.1 depicts the block diagram of the proposed realtime simulator. As shown in Figure 5.1, the two conversing participants are seated in two separate rooms. The two headphones/handsets are connected to the PC running the realtime simulator software via an audio interface. The software incorporates an echo modeling module which reads the data streaming from the audio driver, and imparts the necessary echo path and delay parameters. The processed signal is then routed to the other subject. The simulator software also includes a quality evaluation module, which allows one of the participants (subject in room #1) to control the echo parameters and rate the quality of conversation.

This customized software is executed on a PC running Linux Fedora 16 operating system. As mentioned above, each input signal is recorded, processed, and played back through the audio interface. Within the software, input and output buffering is facilitated by an audio server running as a background process, called the JACK audio connection kit [101].

5.2.1 Audio Interface

The Delta 1010LT soundcard [102] from M-AUDIO was used as the audio interface. This soundcard has eight RCA analog inputs and eight RCA analog outputs and 2 XLR inputs.

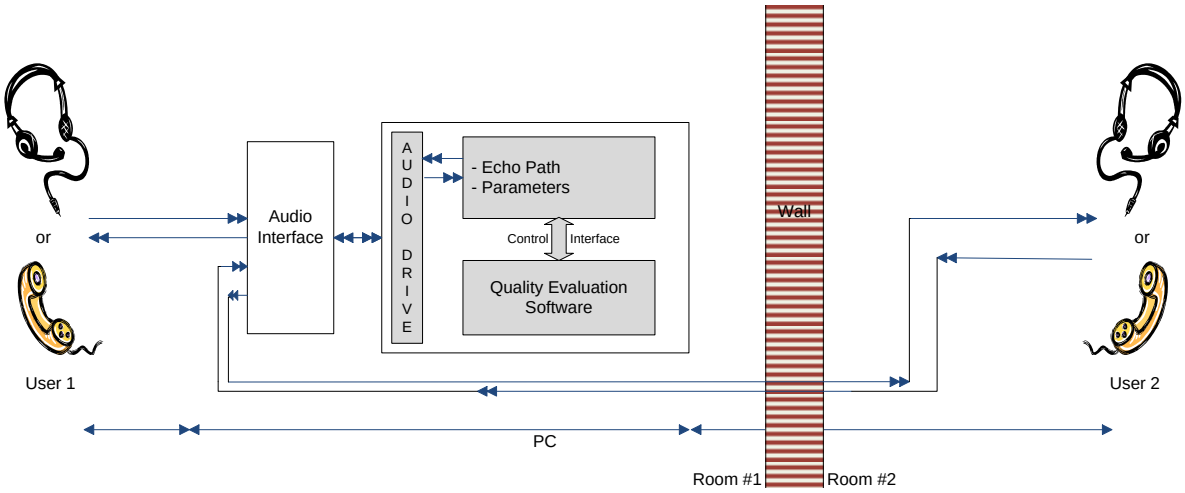

Figure 5.1: Setup for simulating a telephone conversation

An important feature of this sound card that is especially attractive for the present application is its support of sampling rates from 8 kHz to 96 kHz. This allows for the simulation of not only narrowband (8 kHz sampling rate), wideband (16 kHz sampling rate) communication, but also super wideband (32 kHz sampling rate) communication. This scalability with bandwidth is attractive, as the same hardware/software setup can be used to evaluate echo, its parameters, and its mitigation in narrowband, wideband, and super wideband telephony contexts. Another attractive feature of this soundcard is its mixing architecture, which facilitated the generation of the sidetone signal [1]during subjective tests.

5.2.2 Audio Server: JACK

JACK is a low-latency audio server that is especially useful for handling realtime processes [101, 103, 104, 101]. When a program connects to JACK, it gets assigned a number of input and output ports. These ports can be easily connected to each other. JACK also can route these ports to the other applications connected to it, or send them out to the system sound device.

A GUI program, called “Qjackctl”, is used to control the JACK server parameters graphically. This program also provides the status details and error messages generated while JACK is running. Figure 5.2 and Figure 5.3 show the Qjackctl main window and the settings window, respectively. In the following paragraphs, a few parameters that need to be

[1]Sidetone is the signal between mouthpiece and earpiece of the same handset. The presence of this signal enables the users to characterize the circuit as live and help them to adjust the level of their voices.

Figure 5.2: Qjackctl main window

set before running JACK, are briefly explained.

Realtime: when this option is enabled, JACK will run in realtime mode.

Priority: shows the priority that JACK daemon will run in realtime mode. Priority ranges from 0 to 89 and higher value shows that JACK thread is running at higher priority level.

Frames/Period and Periods/Buffer: JACK stores the incoming and outgoing samples in a buffer. Each buffer is divided into transfer units or frames. The length and the number of frames are equal to the values of "Frames/Period" and "Periods/Buffer", respectively. The length of each buffer is the product of the frame length and the number of frames used in each buffer.

Input Device and Output Device: The input device for data capture and the output device for playback are set using these two options.

Input Channels and Output Channels: the number of channels for capture and playback are specified using these options.

Employing JACK interface in an application is simple. The steps listed below were used in our application to interact with the JACK server:
1) Connecting to JACK using the jack_client_open() function.
2) Creating and registering input and/or output ports to enable data movement to and from the program.
3) Registering a process callback which is called by JACK server when there is a need for data processing/transfer.
4) Informing JACK that the program is ready to start processing the data.

Figure 5.4 displays an example connection created with JACK. Here, a client named Mushra was created by the realtime simulator and connected to JACK. This client was then assigned two input and two output ports. The configuration of these virtual ports and their connections to the physical input and output ports of the soundcards is shown in 5.4 as well.

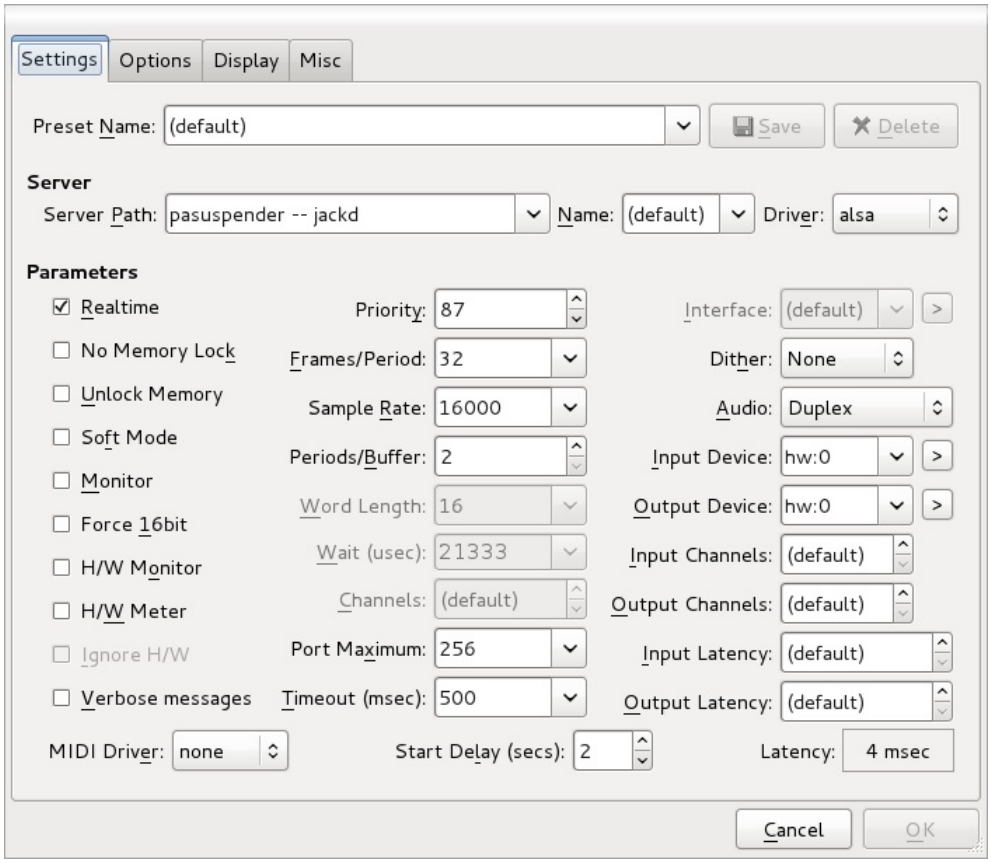

Figure 5.3: Qjackctl setting window

Latency in JACK

Latency can be divided into two modes: input latency and output latency. Input latency is defined as the amount of time taken for input audio signal to be buffered and subsequent generation of an interrupt to enable processing of the buffered data. Output latency is defined as the interval between the time that data is ready for output and the time it takes to fill the output buffer and deliver the analog data.

The minimum latency will be achieved if there are two interrupts per hardware buffer (double buffering). In double buffering, one half of the buffer is used for the audio input and the other half for the audio output. For example, if one half is labeled with 0 and the other half with 1, then as the user application is writing to buffer 0, the hardware driver writes to the buffer 1. Later on, the two buffers are switched to input buffer 0 and output buffer 1.

To summarize, input latency and output latency are determined by the frame length and the buffer size, respectively. Throughput latency is equal to the output latency and is calculated as: $\frac{(buffer\ size)}{(sampling\ rate)}$. Throughput latency is shown at the bottom-right of the settings window

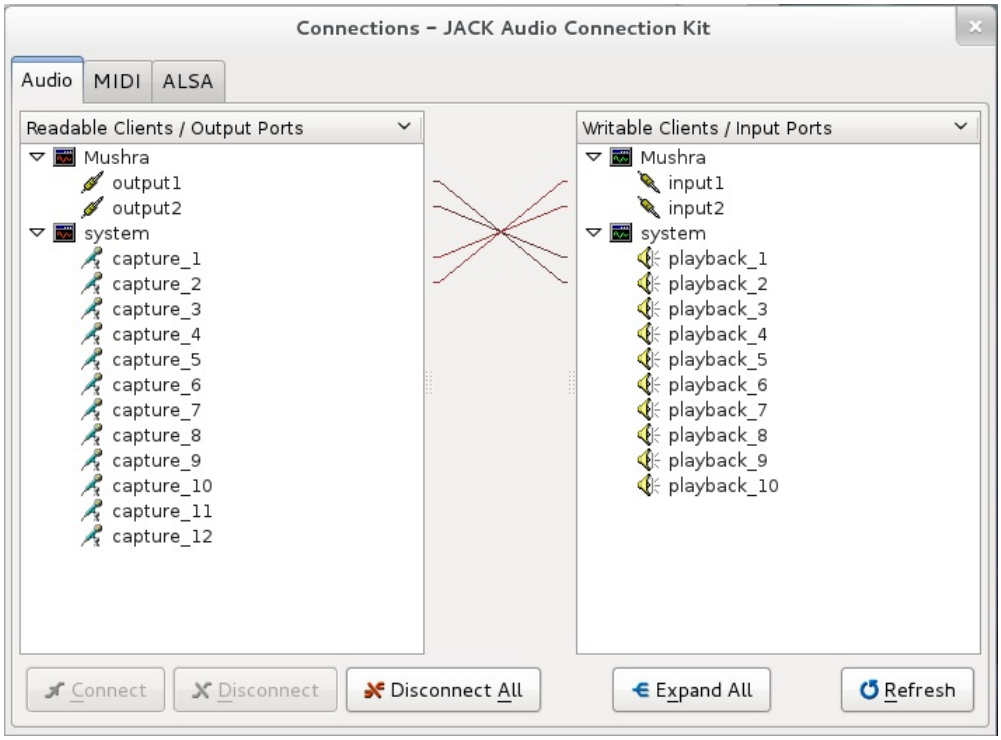

Figure 5.4: Connections window

in Figure 5.3. In this figure, the latency is equal to $\frac{2\times32}{16} = 4ms$.

It is obvious that smaller buffer size results in lower latency and hence smaller delay value. There is, however, one more factor to be considered before determining the buffer size. Whenever JACK is not fed by all the data it requires or when it needs to overwrite the data, an "xrun" (buffer under-runs or over-runs) error occurs. This error occurs when the data does not arrive fast enough and the buffer is not filled in time (under-runs) for the application to use it; or when there is too much data flowing into the buffer and the buffer is overwritten before being processed by the user application (buffer over-runs). A proper buffer size should not cause any of these effects.

5.2.3 Echo Path Simulation Software Module

Figure 5.5 depicts the block diagram of the realtime simulator developed in this thesis for investigating the effect of echo and delay on a realtime conversation. As can be seen in this Figure, this module has two inputs and two outputs, where the two input signals are subject #1's and subject #2's speech signals picked up by the respective microphones, and the two output signals are the signals which are played out at each subject's ears. Since these two

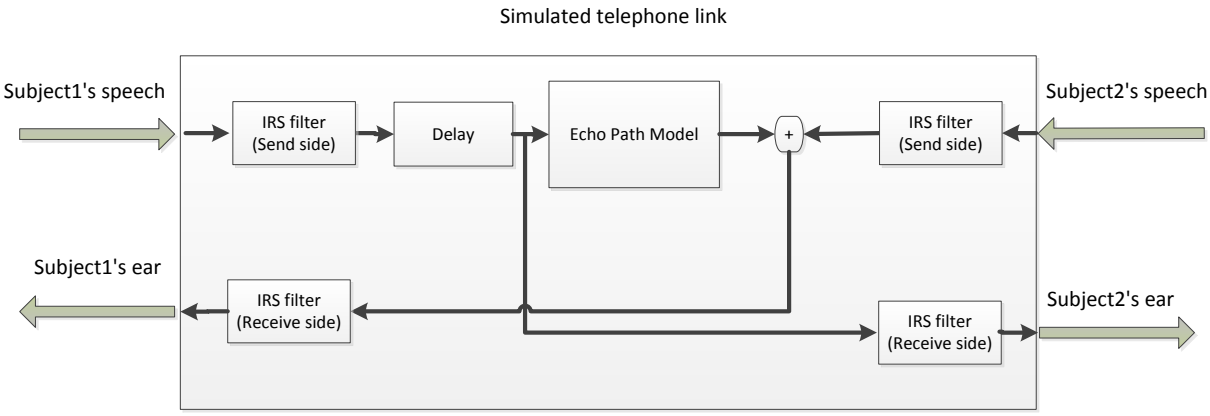

Figure 5.5: Inputs and outputs of the developed simulator

subjects are conversing over the simulated circuit, subject #1's speech should be played out at subject #2's ear and vice versa.

As is evident in Figure 5.5 (and Figure 5.1), only one of the participants will experience the effects of echo. As such, delay and echo are imparted to Subject #1's speech, and it is Subject #1 who rates the quality of conversation.

To simulate the characteristic of the handset, it was necessary to apply intermediate reference system (IRS) filters (send side) on the inputs and (receive side) on the outputs. As the present evaluation concentrated on wideband echo evaluation, the IRS filters should be applicable for wideband application. Since there are a limited number of available wideband handsets, no wideband version of the IRS filters has been standardized by the ITU-T yet. In 2008, Gierlich [19] evaluated the effect of different wideband handset on the perceived speech sound quality. The result of this study has been used in ETSI standard ES 202 739 (V1.3.1) [105] for determining the frequency response masks of the wideband handset or headset.

In 2011, ITU-T P. 311 [106] standard has suggested the upper and lower limits of the sending and receiving frequency characteristics of the wideband handsets; and as a result two target curves have been propounded. The target curves for the send and the receive paths were shown in Figure 5.6 (a) and (b), respectively.

The "delay" block in Figure 5.5 simulates the conversational delay and was realized using a circular buffer, which not only imparted the desired delay but also served as a buffer for the frames received from JACK. The length of the circular buffer therefore depended on

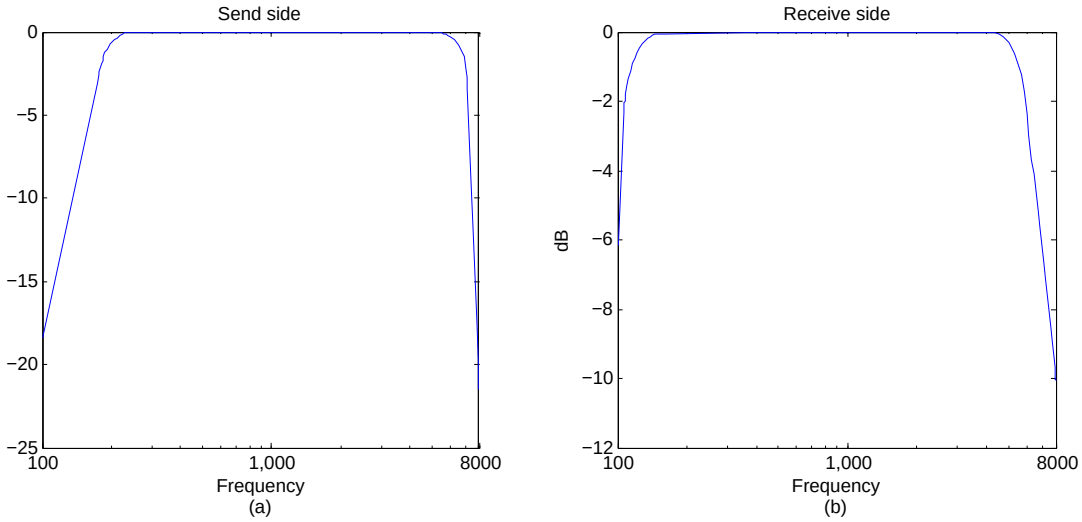

Figure 5.6: Handset (a)sending and (b)receiving sensitivity/frequency mask

the chosen delay value and the length of JACK buffer.

Finite Impulse Response (FIR) filters were used to implement different echo path characteristics. The FIR coefficients representing different types of echo paths are stored as "CSV" files, which are loaded into the program upon selection and applied to the input signal in realtime. The FIR filter output is scaled according to the desired echo attenuation factor.

Realtime implementation of the echo path FIR filter was facilitated by the Intel® Integrated Performance Primitives (IPP) library [107]. Intel IPP is an optimized signal processing library for Intel processors with an array of signal processing functions for realtime computation. The present application used the "ippsFIR" function, which filters a block of consecutive samples through a FIR filter. Prior to the execution of this function, it is essential to allocate memory and initialize the filter state structure which includes the filter taps and the delay line.

With the implementation details of IRS filter, delay and echo path models explained, the next step is to clarify how each frame received from the JACK is processed and sent to the output. The process callback function, includes the following processing: first, the memory addresses associated with the input and output ports are determined; then input1 (subject1's speech) frame is directly written to the memory area that is assigned to the output2 port (subject2's ear). Afterwards, input1's frame is written to the circular buffer and the frame at which the reading pointer is pointing, is read from the circular buffer. This frame is then filtered using the FIR function; the filtered frame then is added to the frame of input2 (subject2's speech) and the result is written to the memory area which has been

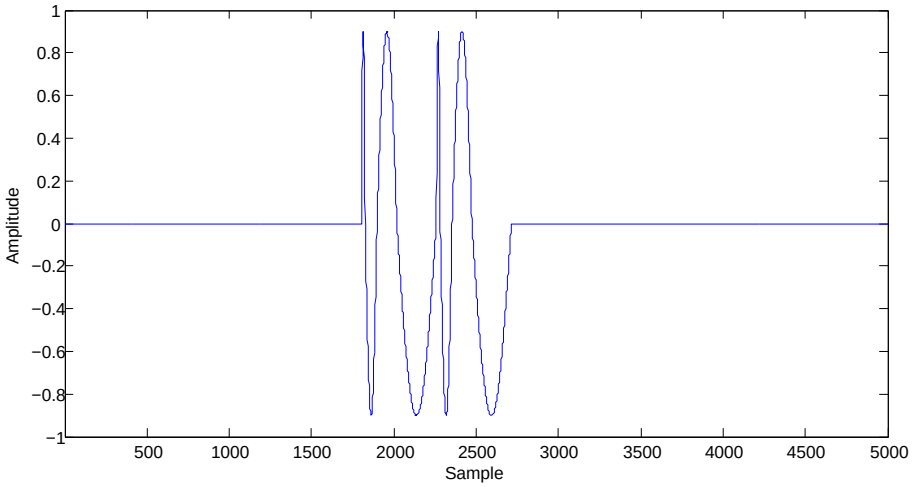

Figure 5.7: The customized waveform used for delay measurement

assigned to the output1 port (subject1's ear).

Total delay

The total delay is cumulative and is determined by the amount of latency introduced by JACK and the processing time of the simulator. In the present work, the sampling rate is set to 16 kHz. The shortest buffer size available was 32 samples which results in a latency equal to: $(32 \times 2)/16 = 4ms$. In order to calculate the total delay, it is required to measure the interval between the time that the signal is picked up from the microphone of subject1 and the time it is sent to subject2's ear.

For this purpose, a customized waveform , shown in Figure 5.7, was generated and the set-up shown in Figure 5.8 was used for the measurement. As shown in this figure, the custom waveform is played out from the laptop and is given to the analog audio input 1 of the PC running JACK and echo path simulator; and both the analog audio input1 and output2 signals of the PC are connected to an oscilloscope. The delay between these two signals, which can be measured on the scope shows the total delay. Figure 5.9 shows a screen shot of the scope we used for our measurement; both input and output signals and the amount of delay between them were shown on the screen. The delay is equal to 13.6 ms, and means that the signal processing being done in our application which includes, writing/reading the signal to/from the circular buffer as well as filtering needs 13.6-4=9.6 ms time to be run.

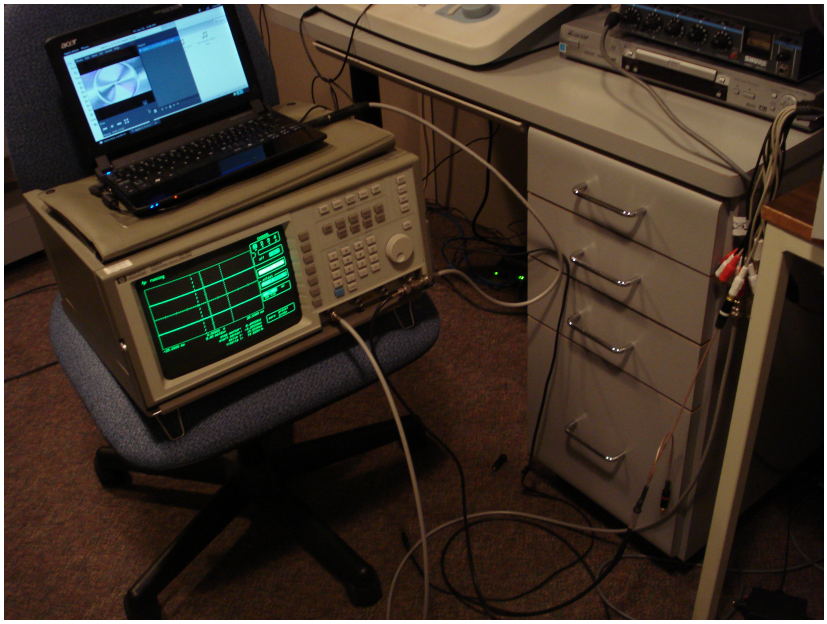

Figure 5.8: The setup to measure the total delay

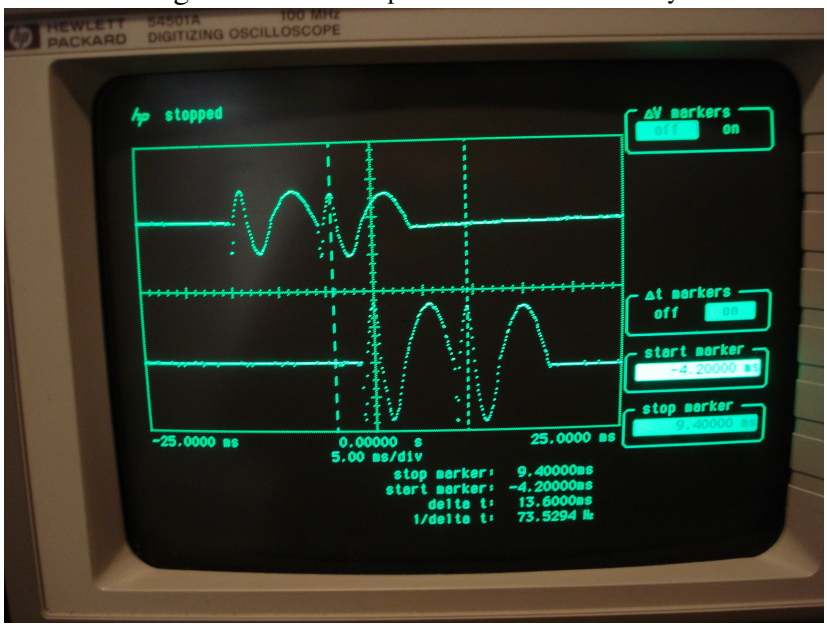

Figure 5.9: A screen shot of the scope

5.2.4 Echo Path Model Normalization

As mentioned earlier, the echo path models are implemented as FIR filters in the software; but these filters need to be normalized before implementation within the software. Two echo path models, one handset mode and one hands-free mode, were simulated in

Figure 5.10: Block diagram for normalizing the echo path model

this research work. For normalization purposes, the reference and echo signal were generated using the procedures shown in Figure 5.10. IRS send- and receive- side filters were applied to the CSS signal (taken from [108]) respectively; the output signal, which was the reference signal, was then filtered by the echo path model to generate the echo signal. The active level of the echo and the reference signals was calculated based on the method presented in [91]. Finally, the scale factor is computed as $Scale\ factor = Active\ level\ of\ the\ \text{“}Reference\ signal\text{”} - Active\ level\ of\ the\ \text{“}Echo\ signal\text{”}$. The normalized echo path model will be : $echo\ path\ model \times \sqrt{10^{(\frac{scale\ factor}{10})}}$

The frequency responses of the normalized echo path models for handset and hands-free at 16 kHz sampling rate are given in Figure 5.11 and Figure 5.12, respectively.

5.3 Subjective Method: Conversation Test

Two double-walled sound booths were used for performing the subjective test. There was one headphone and one microphone within each room. From this point on, room #1 is the room where the subject who rates (subject 1) is seated, and Microphone 1 and Headphone 1 are the ones which were used in this room. The effect of two echo path model, handset and hands-free modes at different test conditions including several delay and ERL values was evaluated. These test conditions are shown in table 5.1. The test conditions were selected based on a report presented by Nortel [79] on user's perception of talker echo, and the descriptions given for each condition in the table show those users 's rating for that condition.

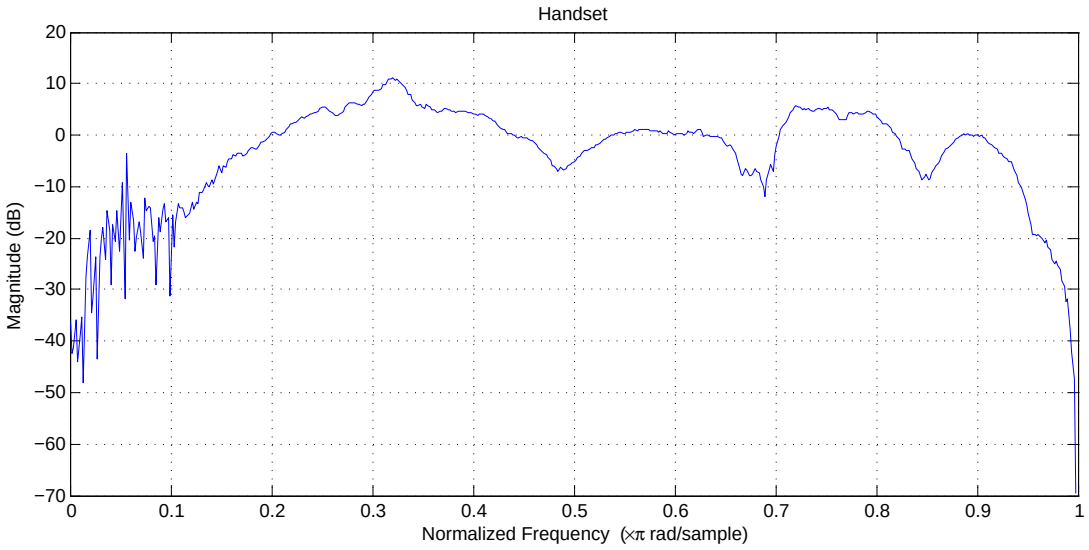

Figure 5.11: The frequency response of the normalized echo path models – Handset mode

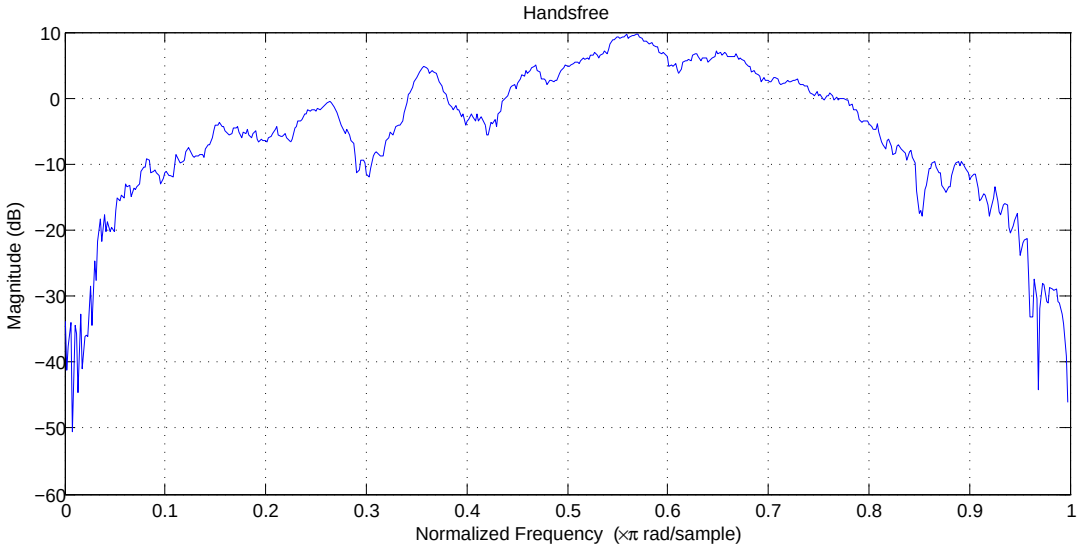

Figure 5.12: The frequency response of the normalized echo path models – Handsfree mode

Table 5.1: Test conditions

	ERL(dB)	Delay (ms)	Description
First screen	80	0	Acceptable audible echo
	40	100	On the border("Acceptable audible echo" and "Unacceptable audible echo")
	20	0	Unacceptable audible echo
	40	200	Unacceptable audible echo
	20	200	Unacceptable audible echo
	10	500	Unacceptable audible echo
Second screen	80	0	Acceptable audible echo
	46	50	Acceptable audible echo
	45	150	Acceptable audible echo
	35	300	Unacceptable audible echo
	20	400	Unacceptable audible echo
	20	600	Unacceptable audible echo

5.3.1 Calibration

To calibrate the microphones and headphones, we first measured the isolation provided by the headphone. Since it was desired to evaluate the effect of delay and echo on conversation, it was important to ensure that no more echo was generated in the other parts such as the headphones. To measure the headphone isolation, a Head and Torso Simulator (HATS) was placed within room 1, where subject 1 will be seated. The mouth simulator of HATS was fed by white noise, and then the sound level was measured at the ear, with the headphone (= 63.10 dB SPL) and without the headphone (= 80.56 dB SPL). So 80.56-63.10 = 17.46 dB SPL isolation was provided by the headphone, which was deemed acceptable.

The microphone 2 was calibrated so that the level of the signal at subject's 1 ear be 70 dB SPL(A). The microphone 2 was fixed at a distance of 10 cm from Mouth Reference Point (MRP) of subject 2.

5.3.2 Subjective Data Collection

Speech material

The subjects were given the scripts for the conversation. The scripts were based on the scenarios of interactive short conversations, taken from typical situations of every day life such as booking a hotel, renting a car, making a reservation in a restaurant and etc.

While subject 2 had the scripts for both sides of the conversation, subject 1 only had his/her own part. There were also some blank spaces in subject 1's script which were to be filled

by this subject based on subject 2's speech during the conversation.

A sample of the conversation script is given in Appendix A. Theses scripts were taken from the websites which provide samples of English language exams.

Participants

Seven pairs of normal hearing listeners were recruited. The participants were young adults at the Western University. None of the participants had explicit knowledge of the simulator and the speech material beforehand. English was the first language for all participants.

Two participants were seated in two double-walled sound booths. There was a XLR microphone and a Sennheiser HDA 200 headphone in each room for the subjects to talk together and hear each other. First, one of the subjects was seated in room 1 and rated the quality and then subjects switched their places and the other subjects was seated in the main room. The experiment lasted about half an hour for each subject.

Test methodology

Internally developed software of the MUSHRA protocol was used for collecting subjective quality ratings. A screen shot of the software is shown in Figure 5.13.

There were 4 different screens that participants worked through and each screen had five play buttons. The five buttons on each screen were randomly associated with different amounts of echo and delay. In fact, there were two screens for each echo path model and the test conditions for each screen are shown in Table. 5.1

The subject who sat in sound booth 1 task was to push one of the play buttons and start to converse with his/her partner by going through one of the blocks of the scripts. The subject can hear his/her partner's voice as well as the echo of his/her own voice; the amount of echo changed in realtime as the subject pushed different play buttons. Subjects were instructed to listen carefully to the echo signal and the partner's speech and indicate the overall quality of the conversation by adjusting the corresponding sliders. The subject was encouraged to switch between different play buttons as many times as they wish, and rate the quality of each with comparing them together. There was also a reference button at the left side of the screen at which no echo and delay is added. Subjects can switch to that if they needed any reference condition.

Since for each condition, the length of the conversation should be long enough to ensure that the subject experiences the effect of echo and delay, it was important that the subject rate the quality after completing the conversation given at least in one of the blocks of the script.

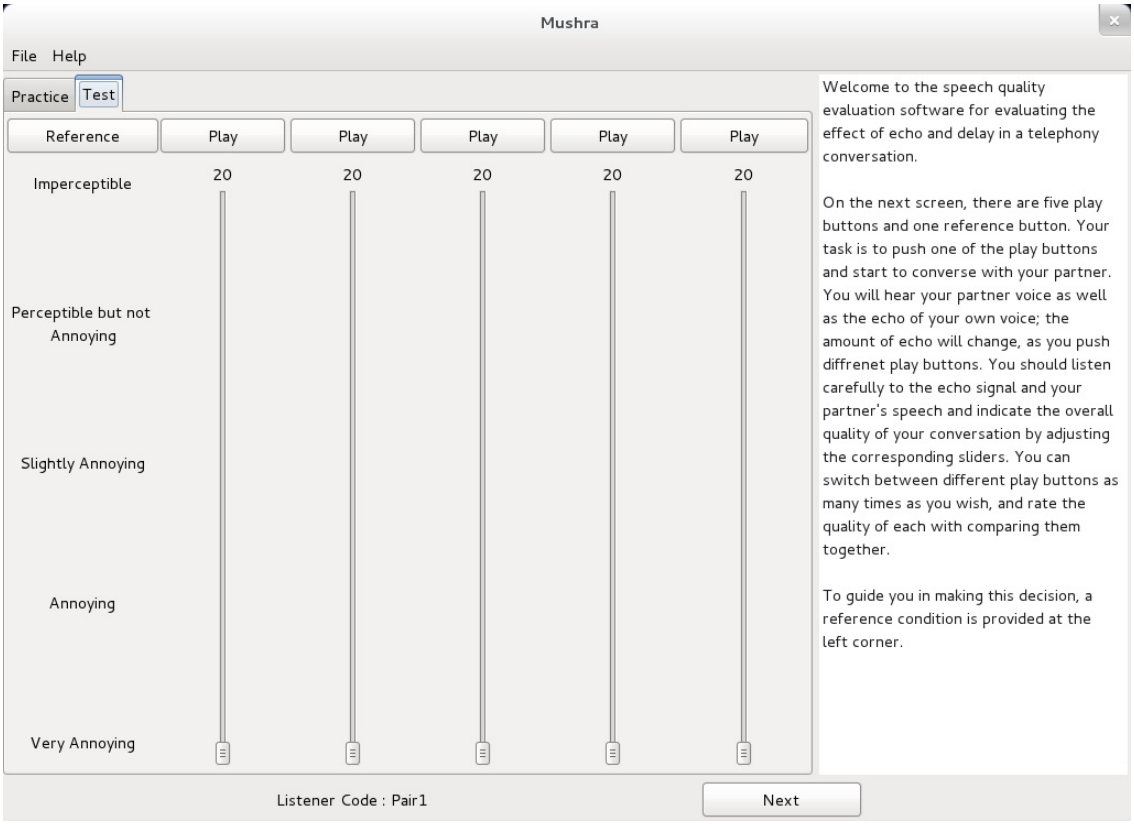

Figure 5.13: Screenshot of the MUSHRA quality ratings software

All four signals, two inputs and two outputs were recorded to the hard disk by the simulator to be later used for developing the objective model. It again shows the importance of having long enough conversation for each condition which results in more data for the model at the given conditions.

After collecting the scores and before analysing them, we noticed that three of the subjects did not rate the quality of a few test conditions which were replaced by the averaged rating scores of the other subject for that condition.

5.4 Subjective and Objective Score Analysis

5.4.1 Reliability of the Ratings

The quality ratings were first analysed for the inter-rater reliability. The consistency of ratings among the 14 participants was measured using Cronbach's α and Intraclass Correlation Coefficient (ICC) using SPSS statistical software package, Version 20.0. For this

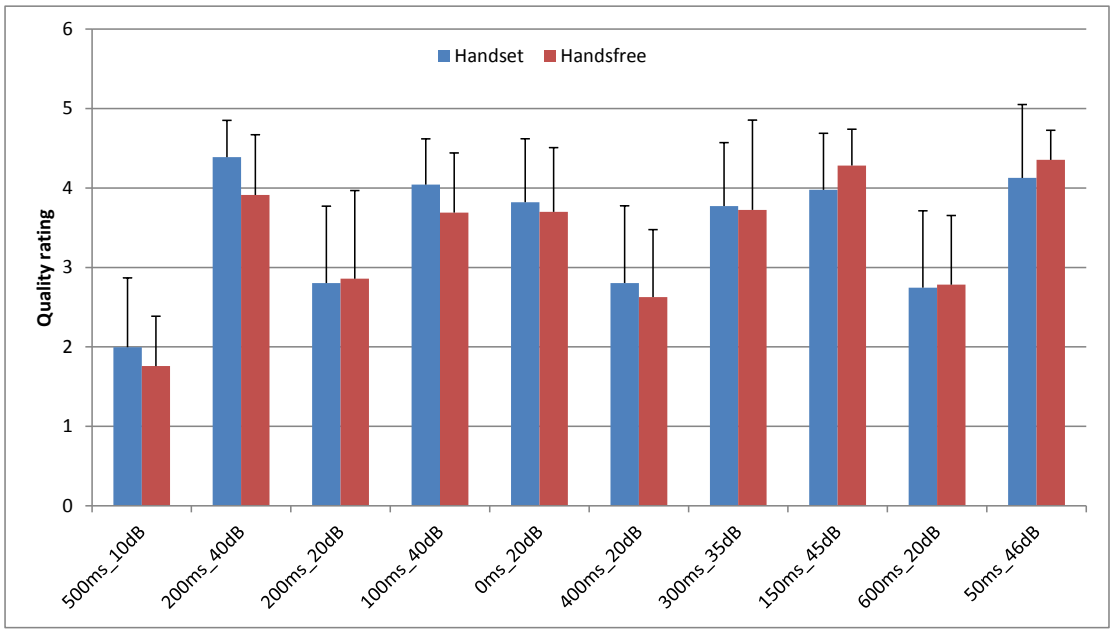

Figure 5.14: Average quality ratings of the conversation across the test conditions

dataset, the Cronbach's α was 0.95 and the average measures ICC ranged between 0.829 - 0.956. These values suggested a very good agreement among listeners on the quality scores.

5.4.2 Averaged Ratings

Figure 5.14 depicts the averaged speech quality ratings for the test conditions, including different amount of delay and echo, in the conversational context. The error bars represent one standard deviation.

It can be seen from the figure that the quality ratings improve with an increase in the ERL value. Furthermore, there were small differences between the two echo path models at some of the test conditions. In order to quantify the significance of these differences, a thorough statistical analysis was performed.

Table 5.2: Results of the post-hoc test

Test conditions	Subset for $\alpha = 0.05$			
	1	2	3	4
'500ms-10dB'	37.54			
'400ms-20dB'		54.02		
'600ms-20dB'		56.28		
'200ms-20dB'		56.46		
'0ms-20dB'			75.50	
'300ms-35dB'			76.09	
'100ms-40dB'			78.35	
'150ms-45dB'			82.95	82.95
'50ms-46dB'			85.45	85.45
'200ms-40dB'				83.63

5.4.3 Statistical Analysis

A split-plot repeated measures ANOVA was first performed in SPSS software with test conditions and echo path models as the "within subject factors". While there was no significant interaction between test conditions and echo path models ($F(3.483, 38.313) = 1.651$ $p > 0.05$), the difference among the test conditions was reported as significant ($F(3.786, 41.649) = 36.702$ $p < 0.05$).

This result shows that the shape of the frequency response of the echo path models did not have a significant effect on the speech quality, and delay and ERL parameters are the ones which had the most effect on the speech quality.

Using the FDR (false discovery rate) control method [73], multiple comparisons were performed to assess the significance of the differences between quality scores obtained from different test conditions. Since there was no interaction between the test conditions and the echo path models, the post hoc test was applied on the quality scores averaged across the echo path models.

Table 5.2 shows the result of this test. The table presents data that will help determine the effect of different ERL and delay values on the speech quality.

As can be observed from this table, the test conditions have been categorised into four main categories. The first category included the test condition ('500ms-10dB') with the minimum ERL value (10 dB). The second category contained test conditions with relatively high delay (200 ms, 400 ms and 600 ms) and low ERL value (20 dB).

Except the condition '100ms-40dB', the test conditions with ERL value equal or higher than 40 dB were grouped in the last subset with the maximum quality ratings. These

results highlight the more significant role of the ERL value in comparison with the delay parameter. Similar result was found for echo quality evaluation in the listening context, discussed in Chapter 3.

On the whole, while we expected to see significant differences between the conditions with same ERL but different delay values, the subjective scores did not reflect this expectation. This result could be because of the way the test was conducted and the effect of delay could have been more prominent if there was more interaction between the subjects or if there were some double talk periods during the conversation.

5.4.4 Objective Quality Evaluation

Using the echo quality subjective scores collected from the conversation test, the performances of LPD and LPD-BMLS in predicting these quality scores were evaluated.

To calculate LPD coefficients, since subject 1 was the subject who experienced the echo and rated the quality of the conversation, the signal at this subject's ear (Out1) was used as the degraded signal and subject 2's speech (In2) as the reference signal. Under the conditions without echo, the subject 2's speech (In2) should be the only signal played out at subject 1's ear (Out1).

As can be seen in Figure 5.5, there is no delay between In2 and Out1, but delay is one of the defining variables for the test and its essence must be captured within the objective model. The effect of delay exists between subject 1's speech (In1) and signal at subject 2's ear (Out2). Using the cross-correlation method presented in [109] the time delay between these two signals was estimated. Comparing the estimated delay with the true delay values shows that, the estimated delay values are about 100 samples (with sampling rate of 16 kHz, it equals to 6.25 ms) more than the true delay values. This additional delay is due the FIR IRS filters in the circuit.

LPD coefficients along with delay parameter were mapped to the true quality scores using the BMLS mapping procedure. Correlation coefficients and standard error of estimation were used for evaluating the performance of LPD alone and LPD+delay-BMLS.

Two types of correlation analysis were used for this purpose. In the first analysis, objective scores were computed for each speech sample which were then used for correlation analysis; i.e. all 271 scores (14 (subjects) $\times$ 2(echo path model)$\times$ 10(test conditions)–9 (note that as mentioned before, three of the subjects missed a few isolated conditions and did not converse over those conditions)) were used for evaluating each objective measure.

Table 5.3: Estimated correlation coefficient and standard error of estimation for LPD and LPD+delay-BMLS for per-sample and condition-averaged analysis.

Measure	**Complete Database**		**Condition-averaged**	
	ρ	σ	ρ	σ
LPD (Total loudness)	0.57	18.55	0.89	7.43
LPD+delay-BMLS (test)	0.66	16.94	0.90	7.01
LPD+delay-BMLS (train)	0.86	11.59	0.98	3.18

In the second analysis, the average score at each condition was used for correlation analysis. In fact, after calculating the objective scores for all speech samples, the scores of all recorded samples at a specific condition (same echo path model and test condition) were averaged. These condition-averaged scores were used for correlation analysis. There were 20 (2(echo path model)$\times$ 10(test conditions)) conditions . As a result, there were 20 pairs of subjective and objective scores for condition-average analysis.

Table 5.3 shows the correlation coefficients and the standard error of estimations for LPD and LPD+delay-BMLS. The results show that both LPD and LPD+delay-BMLS scores resulted in low correlation with the true quality scores for per-sample analysis. The performance of the models did improve when the scores were averaged across the conditions.

It can also be seen that while using BMLS and the delay parameter improves the performance of LPD for per sample analysis ($Z = 2.21 \;\; p < 0.05$), they do not make any significant effects for the condition-average one ($Z = 0.21 \;\; p > 0.05$).

Referring to the subjective score analysis and the comparison between the conditions with different delay values, it can be seen that the effect of delay was not reflected in the subjective quality scores either and the subject did not perceive the effect of the delay as annoying as the effect of ERL value. This may explain why including this parameter as an objective metric did not make any improvement on the quality score prediction.

Figure 5.15 and Figure 5.16 depict the scatter plot of LPD scores and the predicted values by LPD+delay-BMLS respectively, versus the actual scores for (a) per-sample analysis and (b) condition-average analysis. It can be seen from the figures that there was a big variation between the scores for per-sample analysis which was removed by averaging across the conditions.

In summary, even though the methods presented in standards for performing the subjective data collection was followed, it seems that the subjects should have more flexibility to switch between different conditions.

Here, MUSHRA software was used for data collection. In this method subjects are allowed

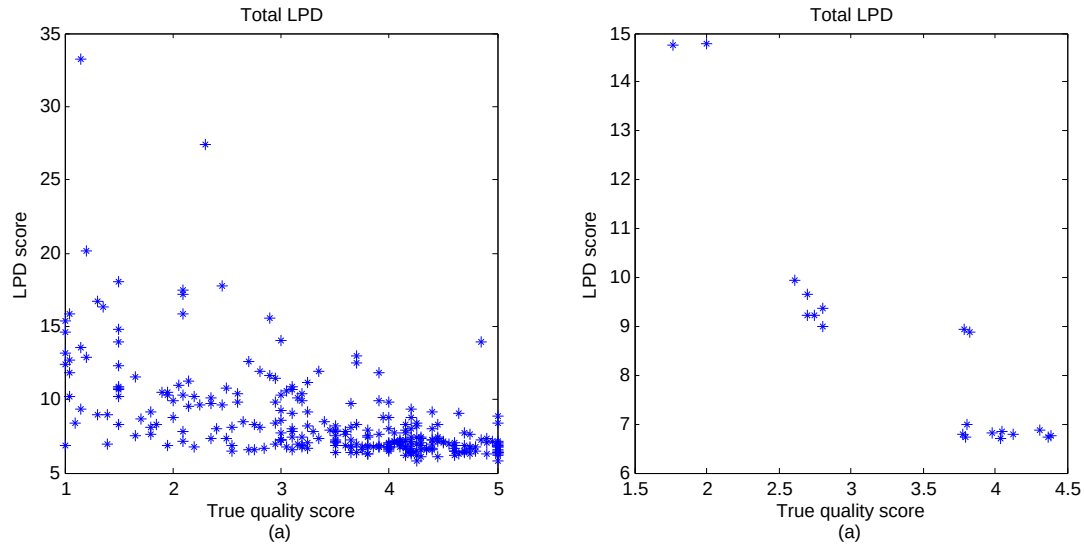

Figure 5.15: Scatter plot of LPD and actual overall quality scores for (a) per-sample analysis (b) condition-average analysis

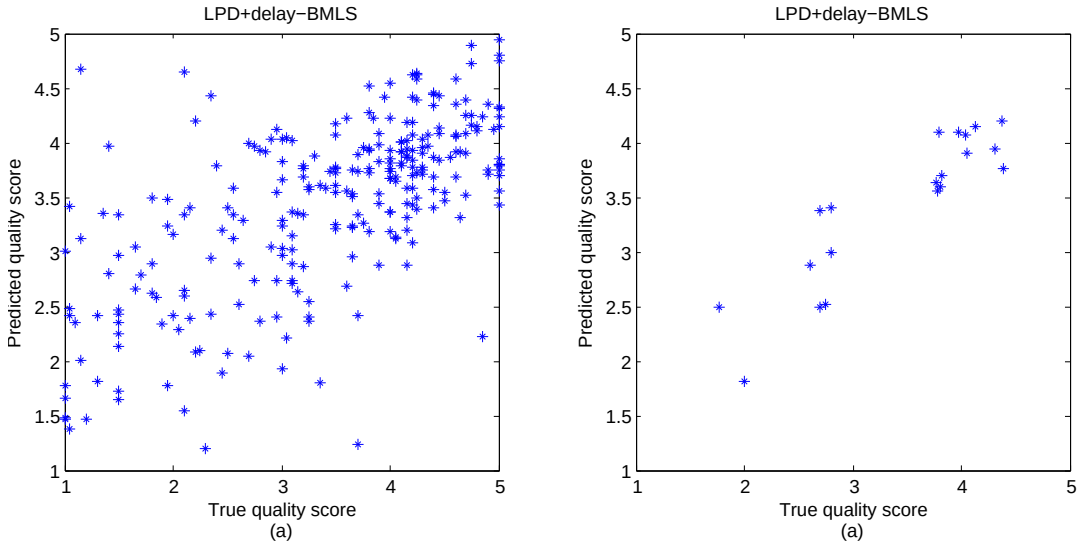

Figure 5.16: Scatter plot of the predicted and actual overall quality scores for (a) per-sample analysis (b) condition-average analysis

to switch back and forth between different test conditions as many times as they want; but since based on the standards the length of the conversation should be long enough for the subjects to experience the effect of the impairments, not many of the subjects switched between the conditions. If less conditions had been provided on each screen or less screens had been designed for each session of the test, the subjects probably would have switched between the conditions more frequently. However, as it was mentioned earlier, all subjective tests especially conversational tests are very expensive and time-consuming.

5.5 Summary

This chapter developed a realtime simulator to evaluate the effect of echo and delay in conversational context. Pairs of subjects conversed over the simulated system and rated the quality of their conversations which were degraded by different amount of echo and delay. The quality scores were analysed. The subjective analysis of the data showed the difference between the test conditions. Using the collected subjective scores, the performance of LPD-BMLS for predicting the conversational speech quality was evaluated.

Chapter 6

Conclusions and Future Work

6.1 Summary

Speech communication over a wider bandwidth (50 - 7000 Hz) has several advantages - improved intelligibility, enhanced speech clarity, and increased user acceptance. However, wideband speech is also vulnerable to environmental and network impairments (eg. noise and echo) over the extended bandwidth. Signal processing algorithms designed to mitigate the effects of impairments must be benchmarked before deployment within the communication system. This thesis focused on one particular aspect of benchmarking *viz.* the effects on perceived speech quality.

Speech quality of a device or algorithm can be assessed through behavioural experiments or through objective, instrumental techniques. While subjective evaluation is the gold-standard, objective speech quality metrics are desired due to economical reasons. However, before the objective speech quality estimates are relied upon, there must be evidence that the estimates can serve as surrogates for subjective quality data. This is usually established through: (a) creating a database of processed speech stimuli, (b) recruiting participants and collecting their ratings of speech quality, (c) computing objective metrics on the same set of speech samples, and (d) correlating the objective metrics with subjective ratings.

This thesis investigated the impact of noise, echo, and their suppression algorithms on wideband speech quality. As databases containing wideband speech corrupted by noise and echo are not readily available, they were created within the present research work. Objective metrics incorporating computational models of human audition and statistical regression functions were developed and validated against subjective data. The following section further details the contributions of this thesis.

6.2 Contributions

- A customized wideband noise reduction database containing speech samples corrupted by three types of background noises at three SNR levels, along with their enhanced versions was created. The overall quality of the speech samples in the database was subsequently rated by a group of listeners with normal hearing capabilities. Comprehensive statistical analyses were performed to assess the reliability of the subjective data, and to assess the performance of noise reduction algorithm across varied noisy conditions. There was a high degree of inter- and intra-subject reliability in the subjective ratings. It was found that the noise reduction algorithms enhanced speech quality for only a subset of the noise conditions.

- The performance of several auditory model-based objective quality metrics, including PESQ, PEMO-Q, HASQI, and LPD was evaluated. While LPD resulted in the best correlation with the subjective scores across the entire database, the metrics performed similarly in predicting speech quality ratings when speech quality scores pertaining to a particular noise condition were averaged.

- For the first time, a particular feature mapping technique, *viz.* Bayesian Multivariate Linear Splines (BMLS), was applied to the problem of wideband speech quality prediction. The BMLS procedure produces a continuous, multi-dimensional, locally linear mapping function for transforming feature vectors into predicted quality scores. A particular advantage of BMLS procedure lies in the automatic adaptation of the order and location of fitted splines based on the data. This is in contrast with other mapping models such as Multivariate Adaptive Regression Splines (MARS), which require explicit model order selection.

- The BMLS mapping procedure was paired with the LPD features (the better performing auditory model based feature vector). Correlational analyses with the wideband noise reduction data showed that the LPD-BMLS had a high degree of correlation with the subjective scores ($\rho = 0.78$ and $\rho = 0.86$ for per-sample and condition-average analysis respectively); while this combination was slightly inferior to the combination of PESQ, WSS, IS and LLR using BMLS and MARS, it is computationally simpler to implement.

- Due to its high degree of correlation with subjective data, LPD-BMLS was used to fine-tune the parameters of the noise reduction algorithms. Another customized

database was developed utilizing the algorithms with the fine-tuned parameters; subjective ratings of the algorithms with the new parameters further validated the performance of LPD-BMLS. A paired comparison test was also performed for comparing the performance of the noise reduction algorithms before and after the updates.

- To investigate the effect of echo in wideband context, a custom database containing speech samples corrupted with different amount of echo and delay and four different acoustic echo path models was created. A listening test was performed and MUSHRA software was used for subjective data collection. The effect of four echo path models including acoustic echo at both narrowband and wideband applications were evaluated. Several end-to-end delay and ERL values were used to generate echo at different levels of annoyance including "not noticeable" echo to "annoying" echo. LPD-BMLS was used for predicting the quality of the speech signals corrupted by the echo. There was a high correlation between the subjective scores and objective scores predicted by these two models. As an application example, the LPD-BMLS model was employed to evaluate the effect of an echo canceller and echo suppressor, and the quality scores predicted by this model followed the expected trend.

- A custom software module was developed to simulate a real world telephone conversation. The realtime simulator was employed for systematically investigating the effect of delay and echo in a conversational context, both subjectively and objectively. Pairs of subjects conversed over the simulated system and only one of them experienced the echo and delay and rated the quality of the conversation. Once again, LPD+BMLS combination was found to be effective in predicting subjective impressions of quality, but only for condition-averaged data.

6.3 Recommendation for Future Work

- The performance of LPD+BMLS can be evaluated in predicting the quality of super wideband speech, where the bandwidth is further extended to 14 kHz with a sampling rate of 32 kHz.

- As mentioned before, the objective metrics evaluated in this thesis are the so-called "intrusive models", which use both reference and degraded signal features for estimating the quality scores. Although ITU has published a standard for a non-intrusive speech quality assessment [110], it is only applicable to narrowband scenarios. Other

non-intrusive metrics such as the speech to reverberation modulation energy ratio (SRMR) [111], and ANIQUE [112] can be investigated as a future research topic.

- As mentioned in the thesis, the Moore-Glasberg (M-G) peripheral auditory model which is used for calculating the LPD metric, is more accurate than Zwicker's auditory model used in the PESQ standard. This is reflected in the better performance of LPD when compared to PESQ for the raw correlation data. Glasberg and Moore [113] presented an extension of this model to applicable to time-varying sounds, and the performance of this time-varying loudness model can be a subject of future research.

- As it was shown, none of the fine-tuned algorithms could improve the quality of speech at all the noisy condition; it seems that there may not be any single parameter set which could make improvements at all the noise types and in different SNR levels. This problem can be solved by using the noise reduction algorithm with different sets of parameters for different types of noises, same as what Choi and Chang [114] did for narrowband noise reduction algorithms.

- Several follow-up projects can be devised utilizing the realtime conversation simulator. For example, the simulator along with the objective quality estimator can be used to evaluate the performance of the echo canceller and echo suppressor algorithms in real world conversational scenarios. Furthermore, the algorithm parameters can be fine-tuned in a manner similar to the procedure described for noise reduction algorithms. Similarly, the simulator can be used to create double-talk impairments to further benchmark the performance of echo canceller and suppressor algorithms.

- Enhancements can be made to the subjective testing of conversational speech quality. Research studies investigating the nuances of interaction (how to induce double-talk?), rating procedures (at what point in the script should the participant rate the quality? How often should they switch between conditions?), and scripts (what scenarios and conversations are better suited for testing?) are warranted.

Bibliography

[1] M. Helfenstein and G. S. Moschytz, *Circuits and Systems for Wireless Communications*. Kluwer Academic Publishers, 2000.

[2] S. Voran, "Listener ratings of speech passbands," in *Speech Coding For Telecommunications Proceeding*, pp. 81–82, 1997.

[3] "7 kHz audio-coding within 64 kbit/s," *ITU-T Rec. G. 722*, 1988.

[4] "Wideband speech coding standards and applications," *VoiceAge White papers*, 2006.

[5] I. Varga, R. D. D. Iacovo, and P. Usai, "Standardization of the AMR Wideband Speech Codec in 3GPP and ITU-T," *IEEE Communications Magazine*, no. May, pp. 66–73, 2006.

[6] S. Pennock and P. Hetherington, "Wideband speech communications for automotive: the good, the bad, and the ugly," in *AES 36th Internation Conference*, (Michigan), pp. 109–116, 2009.

[7] J. Rodman, "The effect of bandwidth on speech intelligibility," *Polycom white paper*, no. September, 2006.

[8] P. Stelmachowicz, A. Pittman, B. Hoover, and D. Lewis, "Effect of stimulus bandwidth on the perception of /s/ in normal- and hearing-impaired children and adults," *Journal of the acoustical society of America*, vol. 110, no. 4, pp. 2183–90, 2001.

[9] D. Macho and Y. M. Cheng, "On the use of wideband signal for noise robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 109–112, 2003.

[10] C. Nadeu and M. Tolos, "Recognition experiments with the speechDat car aurora spanish, database using 8 kHz- and 16kHz-sampled signa1," in *Automatic Speech Recognition and Understanding*, pp. 135–138, 2001.

[11] I. Soon, S. Koh, C. Yeo, and W. Ngo, "Transformation of narrowband speech into wideband speech with aid of zero crossings - rate," *Electronics Letters*, vol. 38, no. 24, pp. 1607–1608, 2002.

[12] U. Heute, "Speech-Transmission Quality: Aspects and Assessment for Wideband vs. Narrowband Signals," in *Advances in Digital Speech Transmission*, p. 572, John Wiley & sons, 2008.

[13] B. C. J. Moore and C.-T. Tan, "Perceived naturalness of spectrally distorted speech and music," *Journal of the acoustical society of America*, vol. 114, pp. 408–419, 2003.

[14] T. A. Ricketts, A. B. Dittberner, and E. E. Johnson, "High frequency amplification and sound quality in listeners with normal through moderate hearing loss," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 160–172, 2008.

[15] R. V. Cox, "Speech Coders: from Idea to Product," *AT&T Technical Journal*, vol. 74, no. 2, 1995.

[16] N. S. Jayant, V. B. Lawrence, and D. P. Prezas, "Coding of speech and wideband audio," *AT & T technical journal*, vol. 69, no. 5, pp. 25–41, 1990.

[17] "Voice quality consumer trial," *Ericsson Consumer & Enterprise Lab*, 2006.

[18] "The E-model - a computational model for use in transmission planning," *ITU-T Rec. G. 107*, 2009.

[19] H. W. Gierlich, S. Poschen, F. Kettler, A. Raake, S. Spors, and M. Geier, "Echo perception in wideband telecommunications scenarios - comparison to E-Model's narrowband echo findings," in *ITU-T workshop on "From Speech to Audio: bandwidth extension, binaural perception"*, (Lannion, France), ITU-T, 2008.

[20] C. Beaugeant, M. Schönle, and I. Varga, "Challenges of 16 kHz in Acoustic Pre- and Post-Processing for Terminals," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 98–104, 2006.

[21] M. Jelinek and R. Salami, "Noise reduction method for wideband speech coding," in *Proc Eusipco*, (Vienna, Austria), pp. 1959–1962, 2004.

[22] J. D. Gordy and R. A. Goubran, "On the perceptual performance limitations of echo cancellers in wideband telephony," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 33–42, Jan. 2006.

[23] B. N. M. Laska, S. Member, and R. A. Goubran, "Improved proportionate subband NLMS for acoustic echo cancellation in changing environments," *IEEE Signal Processing Letters*, vol. 15, pp. 337–340, 2008.

[24] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1124–37, 2008.

[25] E. Larsen, R. M. Aarts, and M. Danessis, "Efficient high-frequency bandwidth extension of music and speech," in *Audio Engineering Society 112th Convention*, (Munich, Germany), 2002.

[26] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.

[27] "Methods for subjective determination of transmission quality," *ITU-T Rec. P. 800*, 1996.

[28] "Mean opinion score (MOS) terminology," *ITU-T Rec. P. 800.1*, June 2006.

[29] "ITU-T coded-speech database," *ITU-T P-series Supplement 23*, 1998.

[30] "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T Rec. P. 835*, 2003.

[31] "Subjective performance evaluation of network echo cancellers," *ITU-T Rec. P. 831*, 1998.

[32] "Head and torso simulator for telephonometry," *ITU-T Rec. P. 58*, 1996.

[33] "Method for the subjective assessment of intermediate quality level of coding systems," *ITU-R Rec. BS.1534-1*, 2003.

[34] G. Stoll and F. Kozamernik, "EBU listening tests on internet audio codecs," *EBU Technical Review*, no. June, 2000.

[35] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–16, 2008.

[36] P. C. Loizou, *Speech enhancement: theory and practice*. CRC Press, 2007.

[37] "Perceptual evaluation of speech quality (PESQ)," *ITU-T Rec. P. 862*, 2001.

[38] "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *ITU-T Rec. P. 862.2*, 2007.

[39] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," *ITU-T Rec. P. 563*, 2004.

[40] "Analysis and interpretation of INMD voice-service measurements," *ITU-T Rec. P. 562*, 2005.

[41] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, 2002.

[42] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *9th International Workshop on Acoustic Echo and Noise Control*, pp. 169–172, Citeseer, 2005.

[43] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Robustness of the hearing aid speech quality index (HASQI)," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.

[44] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 229–238, Jan. 2008.

[45] J. Salmela and V. Mattila, "New intrusive method for the objective quality evaluation of acoustic noise suppression in mobile communications," in *Proc. 116th Audio Eng. Soc. Conv*, 2004.

[46] G. Chen and V. Parsa, "Loudness pattern-based speech quality evaluation using Bayesian modeling and Markov chain Monte Carlo methods," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. EL77–83, 2007.

[47] B. C. J. Moore and B. R. Glasberg, "A revised model of loudness perception applied to cochlear hearing loss," *Hearing reseach*, vol. 188, pp. 70–88, 2004.

[48] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[49] M. Hansen and B. Kollmeier, "Objective modeling of speech quality with a psychoacoustically validated auditory model," *Journal of the Audio Engineering Society*, vol. 48, no. 5, pp. 395–409, 2000.

[50] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception.," *The Journal of the Acoustical Society of America*, vol. 124, pp. 422–38, July 2008.

[51] T. Dau, D. Püschel, and a. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure.," *The Journal of the Acoustical Society of America*, vol. 99, pp. 3615–22, June 1996.

[52] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers.," *The Journal of the Acoustical Society of America*, vol. 102, pp. 2892–905, Nov. 1997.

[53] R. Huber and B. Kollmeier, "PEMO-Q - A new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[54] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, 2010.

[55] B. C. J. Moore and C. T. Tan, "Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion," *Journal of the Audio Engineering Society*, vol. 52, no. 9, pp. 900–914, 2004.

[56] C. C. Holmes and B. K. Mallick, "Bayesian regression with multivariate linear splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 1, pp. 3–17, 2001.

[57] B. Cheng and D. M. Titterington, "Neural networks: a review from a statistical perspective," *Statistical Science*, vol. 9, no. 1, pp. 2–30, 1994.

[58] J. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[59] D. Denison, C. Holmes, B. Mallick, and A. Smith, *Bayesian methods for nonlinear classificaiton and regression*. John Wiley and Sons, 2002.

[60] P. J. Green, "Reversible jump markov chain monte carlo computation and bayesian model Determination," *Biometrika*, vol. 82, pp. 711–732, Dec. 1995.

[61] C. C. Holmes and B. K. Mallick, "Bayesian radial basis functions of variable dimension," *Neural Computation*, vol. 10, pp. 1217–1233, July 1998.

[62] S. Möller, W.-y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

[63] N. Egi, H. Aoki, and A. Takahashi, "Objective quality evaluation method for noise-reduced speech," *IEICE Transactions on Communications*, vol. E91-B, no. 5, pp. 1279–1286, 2008.

[64] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms.," *Speech communication*, vol. 49, pp. 588–601, July 2007.

[65] J. Wang, J. Luo, S. Zhao, and J. Kuang, "Non-intrusive objective speech quality measurement based on GMM and SVR for narrowband and wideband speech," *2008 11th IEEE Singapore International Conference on Communication Systems*, pp. 193–198, Nov. 2008.

[66] S. D. Kamath, *A multi-band spectral subtraction method for speech enhancement*. PhD thesis, Uneversity of Texas at Dallas, 2001.

[67] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 4–7, 1996.

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

[69] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on speech and audio processing*, vol. 11, no. 4, pp. 334–341, 2003.

[70] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[71] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE tra*, vol. 13, no. 5, pp. 857–869, 2005.

[72] P. Kabal, "TSP Speech Database," 2002.

[73] M. Matsunaga, "Familywise error in multiple comparisons: disentangling a knot through a critique of O'Keefe's arguments against alpha adjustment," *Communication Methods and Measures*, vol. 1, pp. 243–265, Dec. 2007.

[74] C. Garbin, "Bivariate correlation comparisons." url http://psych.unl.edu/psycrs/statpage/biv_corr_comp_eg.pdf.

[75] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *Journal of the acoustical society of America*, vol. 100, no. 3, pp. 1703–16, 1996.

[76] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *Journal of the acoustical society of America*, vol. 106, no. 4, pp. 2040–50, 1999.

[77] M. Hansen, "Effects of multi-channel compression time constants on subjectively perceived sound quality and speech intelligibility," *Ear and hearing*, vol. 23, pp. 369–80, Aug. 2002.

[78] A. Perry, *Fundamentals of voice-quality engineering in wireless networks*. Cambridge University Press, 2007.

[79] M. Armstrong and F. Blouin, "Echo canceller engineering guidelines," *Nortel Networks*, 2005.

[80] J. W. Emling and D. Mitchell, "The effects of time delay and echoes on telephone conversations," *The Bell System Technical Journal*, pp. 2869–2891, 1963.

[81] "Methodology for derivation of equipment impairment factors from subjective listening-only tests," *ITU-T Rec. P. 833*, 2001.

[82] L. Ding, M. S. El-hennawey, and R. A. Goubran, "Nonintrusive measurement of echo-path parameters in VoIP environments," *IEEE transactions on Instrumentation and Measurement*, vol. 55, no. 6, pp. 2062–2071, 2006.

[83] R. A. Sukkar, "Echo detection and delay estimation using a pattern recogntion approach and cepstral correlation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 909–912, 2007.

[84] L. O. Nunes, F. R. Ávila, A. F. Tygel, L. W. P. Biscainho, B. Lee, A. Said, and R. W. Schafer, "A Parametric objective quality assessment tool for speech signals degraded by acoustic echo," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 8, pp. 2181–2190, 2012.

[85] "Method for objective measurements of perceived audio quality," *ITU-R Rec. BS.1387*, 2001.

[86] N. Pourmand, D. Suelzle, V. Parsa, Y. Hu, and P. Loizou, "On the use of bayesian modeling for predicting noise reduction performance," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3873–3876, 2009.

[87] L. W. P. Biscainho, P. a. a. Esquef, F. P. Freeland, L. O. Nunes, a. F. Tygel, B. Lee, A. Said, T. Kalker, and R. W. Schafer, "An objective method for quality assessment of ultra-wideband speech corrupted by echo," *IEEE International Workshop on Multimedia Signal Processing*, pp. 1–6, Oct. 2009.

[88] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. on Digital Signal Processing*, pp. 1–4, 2009.

[89] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony ?," in *20th International Congress on Acoustics (ICA)*, no. August, (Sydney, Australia), 2010.

[90] "Digital network echo cancellers," *ITU-T Rec. G. 168.*, 2009.

[91] "Objective measurement of active speech level," *ITU-T Rec. P. 56*, 1993.

[92] J. Benesty, T. Gänsler, D. Morgan, M. Sondhi, and S. Gay, *Advances in Network and Acoustic Echo Cancellation*. Springer, 2001.

[93] I. Cohen, J. Benesty, and S. Gannot, *Speech Processing in Modern Communication: Challenges and Perspectives*. Springer, 2010.

[94] E. R. Ferrara, "Fast implementation of LMS adaptive filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, 1980.

[95] J.-S. Soo and K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[96] L. a. Thorpe, "Subjective evaluation of speech compression codecs and other non-linear voice-path devices for telephony applications," *International Journal of Speech Technology*, vol. 2, pp. 273–288, May 1999.

[97] R. R. Riesz and E. T. Klemmer, "Subjective evaluation of delay and echo suppressors in telephone communications," *The Bell System Technical Journal*, vol. 42, pp. 2919–2941, 1963.

[98] E. T. Klemmer, "Subjective evaluation of transmission delay in telephone conversations," *The Bell System Technical Journal*, pp. 1141–1147, 1967.

[99] G. K. Helder, "Customer evaluation of telephone circuits with delay," *The Bell System Technical Journal*, vol. 45, p. 11571191, 1966.

[100] G. Williams and L. Moye, "Subjective evaluation of unsuppressed echo in simulated long-delay telephone communications," *Proceedings of the Institution of Electrical Engineers*, vol. 118, no. 3-4, pp. 401–408, 1971.

[101] "JACK Audio Connection Kit." url http://jackaudio.org/.

[102] "Delta 1010LT." url http://www.m-audio.ca/products/en_ca/Delta1010LT.html.

[103] Davis Paul, "JACK-AUDIO-CONNECTION-KIT," in *Linux Audio Developers*, (Karlsruhe), 2003.

[104] R. Hallum, "JACK." url http://www.audiosite.org/uploads/4/3/4/3/4343108/inter-appn-audio-osx-3.pdf, 2008.

[105] ETSI, "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP terminals (handset and headset) from a QoS perspective as perceived by the user," 2009.

[106] "Transmission characteristics for wideband digital handset and headset telephones," *ITU-T Rec. P. 311*, 2011.

[107] Intel, "Intel ® Performance Primitives for Intel ® Architecture ," *Signal Processing*, vol. 1:Signal P, no. March, 2009.

[108] "Test signals for use in telephonometry," *ITU-T Rec. P. 501*.

[109] S. Voran, "Objective estimation of perceived speech quality - Part I : development of the measuring normalizing block technique," *IEEE transaction on speech and audio processing*, vol. 7, no. 4, pp. 371–382, 1999.

[110] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," *ITU-T Rec. P. 563*, 2004.

[111] T. H. Falk and W. Y. Chan, "A non-intrusive quality measure of dereverberated speech," in *Internation warkshop for acoustice echo and noise control*, 2008.

[112] D.-s. Kim, "ANIQUE: an auditory model for single-ended speech quality estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 821–831, Sept. 2005.

[113] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[114] J.-H. Choi and J.-H. Chang, "On using acoustic environment classification for statistical model-based speech enhancement," *Speech Communication*, vol. 54, pp. 477–490, Mar. 2012.

Appendix A

You are subject A. Please start the conversation.

Practice

A	**A:** Hi Julie, Pat here from HR. Could I please check a few details you've put on your July expenses form? **A:** Firstly, you've written $300 for entertainment, but you haven't said what it was for - I presume it was something to do with the visitors from that engineering firm in Margate. **Which group**: ... **A:** OK. That's all. Can you submit your receipts to the finance division as soon as you can? Christine can then process your claim.
B	**A:** Morning. It's John here. Do you have a few minutes to talk about next month's edition of "Business News"? **A:** No, not this edition. I'd like to look at the area of insurance. In fact, I've just set up an interview with Martin Drew. **A:** I think his title's actually Financial Controller, but, yes, he's the one.
C	**A:** That's right - in the customer services section. **A:** I thought it'd be better to look at where the company's going, about its new London offices, new staff, that kind of thing. **A:** I will, thanks!

1- Shipping

1	**A:** Good morning Packham's Shipping Agents. Can I help you? **Enquiries about ...** . **A:** Where do you want to send your box Ma'am? **from ... to ...** . **A:** Yes, of course. Would you like me to try and find some quotations for you? **A:** Well first of all, I need a few details from you. Can I take your name? **Name: ...** .
2	**A:** Thank you, and you say that you will be sending the box to Kenya? **A:** And where would you like the box picked up from? **From: ...** . **A:** Yes, of course. I'll take down the address now. ... . **A:** And where's that? **Address: ...** .
3	**A:** Oh yes, I know it. And the postal code please? **Postal code: ...** . **A:** Right ... and I need to know the size **Size: ...** . **A:** Right. **Size: ...** . **A:** Ok **Size: ...** . **A:** Great. So I'll calculate the volume in a moment and get some quotes for that. But first can you tell me, you know, very generally, what will be in the box? **Inside the box: ...** .

4	**A:** OK. Good. Anything else? ... . **A:** OK, and what is the total value of the contents? **Worth: ...** . **A:** Ok, let me see how much it costs for you. It would be $250 for your box. **A:** How do you want to pay for that? ... . **A:** That's fine. Can I have your visa card number please? ... .
5	**A:** And when is its expiry date please? ... . **A:** Could you please give me your contact number please? ... . **A:** We can pick up your box tomorrow at 3 p.m., does it work for you? **When: ...** . **A:** That's fine. Is there anything else I can help you with? **A:** Thanks for choosing our agency, and have a good day.

2- Rent a Van

6	**A:** Hello, this is Munster Car and Van Hire, how can I help you? **Reason of the call: ...** . **A:** Certainly, Ma'am. We have a number of vehicles in stock right now. Though I should warn you that it's our busy time of year and you'll need to book today if you want to guarantee a vehicle. What precisely are you looking for? **What type: ...** . **A:** Well. Let me see. We have the latest Renault which has the largest capacity and then there is the Ford which is our most popular hire. **What to carry: ...** .
7	**A:** Oh yes. I don't see any real problem with that. You could fit a good 18 cubic metres in there. **Which car : ...** . **A:** You mentioned next week. Can you give me a precise date so that I can go ahead and make the booking for you? **When: ...** . **A:** It's $21 a day and $12 a half day so that would be $33 extra. **A:** And just to be clear, you intend to return it Monday morning. **Return on: ...** .
8	**A:** Now I'll need to take some details from you so that I can fill out all the paperwork. Your name is? **Name: ...** . **A:** Okay. And could you give me your address please? **A:** We'll take the old address then. We just need somewhere where we can send the confirmation of your booking to. **Address: ...** . **A:** Next thing on the list is a telephone number. Given that you're in between houses. I think it makes sense to take your cell phone number. We need a means to contact you in case of emergency. **Number: ...** .

9	**A:** Now I need some details about your driving record for insurance purposes. I take it you have a full driving licence and not a provisional one. **A:** No. no. Anyone with a full driving licence is qualified to drive a van of that size. Any convictions? **Any convictions: ...** . **A:** Okay. They don't count. So that is no convictions. Have you had any insurance claims in the past 5 years? **Insurance claim: ...** .
10	**A:** Almost there now. You'll need to show some ID when you collect the van. Anything with your photo on it will do just fine. **Type of ID: ...** . **A:** And how will you be paying Ma'am? You don't have an account with us, do you? We take all the major credit cards. **Type of card: ...** . **A:** Can you give me your card number. **Card number: ...** . **A:** Could you give me its expiry date please? **Expiry date: ...** . **A:** Thanks Ma'am. See you on Friday.

3- Business courses

11	**A:** Hello, this is Business Nationwide, Daniel speaking, how can I help you? **Reason of the call: ...** . **A:** We offer two courses which may be of interest to you. Our first course is called 'Getting Started'. It's a two-hour evening course, and it runs from 6pm to 8pm. We discuss things like "Is starting a business right for me?", writing a business plan and some of the legal issues. It runs at various locations in the area. Where are you based? **Living place: ...** . **A:** Eastleigh. So, the closest course to you would be in Handbridge, and the next one is on the 20th March. **Question about: ...** .
12	**A:** we discuss the following topics: "Creating a business case action plan", "Initiating a business case action plan", "Developing the initiative", "Analyzing your options", "Managing risk and Ranking alternative solutions". **A:** That one is free. Do you want to participate in that? **Question about: ...** . **A:** This course is appropriate for anyone who wants to develop a clear business case for decision-making. There are no prerequisites.
13	**A:** But did you say you're trading already? When did you start your own business? **When: ...** . **A:** Well, you might be better off taking our three-day course - 'Business Basics'. It's not free I'm afraid. It's subsided and costs $80 for the three days, unless you've been unemployed in the past six months, in which case it's just $20. **A:** Well, it's well worth the money. The three days cover the essential aspects of running a business. The first day covers legal issues, such as tax, insurance, employment laws and health and safety. The second day covers marketing and pricing, and the third covers accounting and book-keeping.

14	**A:** Let me see. No, it's not available in Handbridge, I'm afraid. The nearest course to you would be in Renton. There's one on the 5th March, and another on the 18th April. **A:** I'll send out a pack to you if you like, with some details of the courses and also some information about what you need to do to set up and who you need to register with. **A:** Can I take your name? **Name: ...** .
15	**A:** And can you give me your address please? **Address: ...** . **A:** And have you got an email address? If so, we can send you details of any courses that are happening near you that you might be interested in. **Email address: ...** . **A:** Great! well, I'll have the information pack sent out to you today. **A:** My pleasure. Bye.

4- Adventure

16	**A:** Good morning, White Cloud Adventures. How can I help you? **Reason of the call ...** . **A:** Okay, well we offer a number of tours. Where were you thinking of going? **Which trip: ...** . **A:** Yes, that's our Dolphin Watch Explorer. It's very popular so you usually need to book in advance. **A:** It's actually 1.5 days and that includes the two hour travel time each way from here.
17	**A:** Yes, you leave from here, but you travel by coach, then split into smaller groups for the boat trip. There are up to 20 people on each boat, though we run tours for numbers over 15. The boats travel to different spots in the ocean, so we don't have too many people disturbing the dolphins. **A:** Well, because its peak time now there is a tour every day, though we run them once a month outside of peak season and at certain times of year we close the tours down completely for a while. **Question about: ...** .
18	**A:** Just let me check on the computer. Okay well the next available booking is six days from now on (Q3) December 18th and then we don't have any free places on another tour until December 23rd. **Question about: ...** . **A:** That's actually quite unusual, though of course the dolphins can be unpredictable! Actually this is a very good time of year to see them and we have a policy that if you don't get to see any throughout the 1 1/2 days you get a 50% discount on a second booking.

19	**A:** Yes, and you get to stay in really nice accommodation too. I would say the majority of our customers rate the tour really highly. The hotel we use is called the Kaoriland Lodge, it's right on the beachfront and has a lovely restaurant and gardens. **A:** Almost everything! The price includes transportation to and from the hotel, the boat trip, overnight accommodation, lunch on the boat and breakfast at the hotel the next day. You do need to pay for your evening meal though, that isn't included in the price. The hotel restaurant is quite reasonably priced and there are one or two nice eating places just a few minutes walk away too, if you feel like exploring. **Question about: ...** .
20	**A:** Well drinks and lunch are included in the package, so you really just need to take your swimwear, a towel and some warm clothing as it can get quite windy on the boat. Don't forget your camera and some sun screen. Oh, and I forgot to mention you can hire snorkels and kayaks very cheaply while you're on the trip, so you can have fun in the water when the boat stops for lunch. **A:** Well, you'll get back to the hotel at around 5.30pm, it has lovely hot pools and there is also a sauna and gym. You have to pay extra to use those. But there is also a floodlit tennis court and indoor swimming pool which are free to guests. **Question about: ...** .
21	**A:** It is usually $420 at this time of year; but there is a discount for bookings of more than 3 people which brings the price down to $390 per person. **How many: ...** . **A:** Just let me check, I think that should be okay. ... Yes, that's fine. We have 6 places left at the moment. **A:** Alright. Well, you need to pay a 50% deposit at the time of booking. That is fully refundable if for any reason you have to cancel more than 48 hours before departure date. The full amount must be paid at least 2 days before departure. If you cancel within 48 hours of the tour, you have to pay 25% cancellation fee. **Type of card: ...** .
22	**A:** Yes, credit card is fine. So I'll make a booking for four people for the 18th of December. You'll need to be here at this office by no later than 7.45am on the day of departure. The coach leaves promptly at 8am. **A:** Well that's all booked for you now. Write down this booking number too, just in case you lose your tickets or anything. **A:** It's 2-6-5-9-1-T. That's T for tango. That's all done for you. Hope you have a great trip!

5- Book a hotel

23	**A:** Worldbridges Travel Agency. Good morning. Can I help you? **How many: ...** . **Where: ...** . **A:** Quite difficult in July, would you like to stay at a hotel? Or are you thinking of a villa or an apartment? **What accommodation : ...** . **A:** Provided it's not during July, yes. You know prices are lower out of season. How long would you like to stay? **How many days: ...** . **Budget: ...** .
24	**A:** For that price you won't have many options, I'm afraid, but let me find out. If you could arrange to make it in late June. I might have a bedsitter for $75. It could accommodate 3 single beds, and it's 5 minutes' walk from the main beach in Mykonos. **A:** Well, contact your friends, come to an agreement and give me a ring again. My name is Arnold Smith, you'll find me here any working day from 10 am to 6 pm, but not on Saturdays. Remember we only have a month left, so you need to make up your minds I'd say today or tomorrow! **A:** Wait ! You haven't given me your name. **Name: ...** . **A:** Thanks.

6- Book a table in a restaurant

25	**A:** Hello, Fusion Restaurant. How can I help you? **Reason of call: ...** . **A:** Yes, we do. **Number: ...** . **A:** When is it for? **When ...** . **A:** The 16th of November? **A:** OK, well, usually we offer a set menu for groups up to 25 people. If you want, you can order a la carte, but it usually works out more expensive and obviously if everybody orders different food, then it can take us a lot longer to prepare. And particularly on a Friday as we're usually pretty busy.
26	**A:** Obviously it's up to you, but it's a good option and you get to try different types of food. **Diets: ...** . **A:** That's not a problem. We cater for all dietary requirements. We offer several different menus and you can choose which one you prefer. Our basic menu costs $25 per person. We do it for a minimum of 4 people, as with all our menus. That's menu A. Then they each go up in price depending on which one you order. Menu B is $30 per person. Menu G is $35 per person and menu J is $40 per person.
27	**A:** Well, all our set menus include a variety of food - some Asian style, some Mediterranean, some Latin American and some British food. **Question: ...** . **A:** Yes, that's right. Basically you get a selection of starters for the whole table, for example in menu A, you get a goat's cheese salad, onion bhajis and guacamole. With menu J, you get spare ribs, king prawns, hummus and pitta bread and a selection of salads. You get more options with the more expensive menus.

28	**A:** You have a choice of 3 options. All menus also include a dessert and coffee. **A:** Of course. Or I could email them to you, if you want, to save you the trouble. **A:** So, what's your name? **Name: ...** . **A:** OK. And your e-mail? **Email: ...** . **Phone ...** . **A:** All right, I will make a temporary reservation for you? how many people did you say? **Number: ...** . **A:** OK, that's fine. You can confirm numbers and which menus you want a couple of days before and I'll send you the email with the menus right now.

7- Confirming a hotel booking

29	**A:** Good afternoon, Orion Hotel, how may I help you? **Reason: ...** . **A:** I'm sorry the line is rather bad, would you mind repeating that, please? **Name: ...** . **A:** And when was the reservation for? **When: ...** . **A:** Let me just check if we have your details on the system. Yes, here we are. **A:** I'm just getting your booking details up now. Yes, we do have another double available for those nights. Was there anything else? **What else: ...** .
30	**A:** Yes, we do, but by prior arrangement. But now you've requested it, I'll put that down in the booking. How many people was that for? **How many: ...** . **A:** Very good. Anything else? **A:** No problem sir, So you'd like two double rooms for the nights of 23rd to the 29th July inclusive, vegetarian provision for one and an early morning call on your departure. Is that correct? **A:** If we can help you with anything else, just give us a ring. We look forward to seeing you in July.

You are subject B (Please wait for subject A to start the conversation)

Practice

A	A: Hi Julie, Pat here from HR. Could I please check a few details you've put on your July expenses form? **B:** Of course, go ahead. A: Firstly, you've written $300 for entertainment, but you haven't said what it was for - I presume it was something to do with the visitors from that engineering firm in Margate. **B:** No, that was later. This was a group of our salespeople that have recently been taken on at the London office. A: OK. That's all. Can you submit your receipts to the finance division as soon as you can? Christine can then process your claim. **B:**OK, thanks.
B	A: Morning. It's John here. Do you have a few minutes to talk about next month's edition of "Business News"? **B:** Sure. I was going to call you actually, to check you still want it to be about business trends. A: No, not this edition. I'd like to look at the area of insurance. In fact, I've just set up an interview with Martin Drew. **B:** You mean, the new Financial Manager at Mannifold. A: I think his title's actually Financial Controller, but, yes, he's the one. **B:** Hasn't the company just won the Davy Business Prize?
C	A: That's right - in the customer services section. **B:** So, what is going to be the focus of the article âĂŞ reasons for the company's success? A: I thought it'd be better to look at where the company's going, about its new London offices, new staff, that kind of thing. If you need any help, just let me know. A: I will, thanks!

1- Shipping

1	A: Good morning Packham's Shipping Agents. Can I help you? **B:** Oh yes, I'm ringing to make enquiries about sending a large box. A: Where do you want to send your box Ma'am? Back home to Kenya from the UK A: Yes, of course. Would you like me to try and find some quotations for you? Yes, that'd be great. A: Well first of all, I need a few details from you. Can I take your name? **B:** It's Rachel Donald.
2	A: Thank you, and you say that you will be sending the box to Kenya? **B:** That's right. A: And where would you like the box picked up from? **B:** From college, if possible. A: Yes, of course. I'll take down the address now. **B:** It's Westall College. A: And where's that? **B:** It's Downlands Road, in Bristol.
3	A: Oh yes, I know it. And the postal code please? **B:** It's B8S 9P5 A: Right ... and I need to know the size **B:** Yes, it's 1.5m long. A: Right. **B:** 0.75m wide A: Ok **B:** And it's 0.5m high or deep. A: Great. So I'll calculate the volume in a moment and get some quotes for that. But first can you tell me, you know, very generally, what will be in the box? **B:** Yes, there's mostly clothes and some books.

4	A: OK. Good. Anything else? **B:** Yes, there are also some toys. A: OK, and what is the total value of the contents? **B:** It is about $1700. A: Ok, let me see how much it costs for you. It would be $250 for your box. **B:** That's fine. A: How do you want to pay for that? **B:** By visa card, if is it OK? A: That's fine. Can I have your visa card number please? **B:** Yes, it is 5-6-4-3-2-5-7-4-8-6-8-0-9-7-6-4
5	A: And when is its expiry date please? **B:** It is on September, 2015. A: Could you please give me your contact number please? **B:** Its 546-543-7687 A: We can pick up your box tomorrow at 3 p.m., does it work for you? **B:** Oh, No. I have an exam at that time. What about Friday at 3 pm. A: That's fine. Is there anything else I can help you with? **B:** That's all. Thanks A: Thanks for choosing our agency, and have a good day. **B:** Thanks. You too

2- Rent a Van

6	A: Hello, this is Munster Car and Van Hire, how can I help you? **B:** I was wondering whether you have a van available for hire next week. A: Certainly, Ma'am. We have a number of vehicles in stock right now. Though I should warn you that it's our busy time of year and you'll need to book today if you want to guarantee a vehicle. What precisely are you looking for? **B:** Well. I'm looking for a small van. A: Well. Let me see. We have the latest Renault which has the largest capacity and then there is the Ford which is our most popular hire. **B:** Would the Ford be large enough to take my garden table and chairs?
7	A: Oh yes. I don't see any real problem with that. You could fit a good 18 cubic metres in there. **B:** Okay then the Ford it is. A: You mentioned next week. Can you give me a precise date so that I can go ahead and make the booking for you? **B:** I'd like to pick it up on Friday and then return it first thing on Monday, A: It's $21 a day and $12 a half day so that would be $33 extra. **B:** That's not so bad - cheaper than I expected. A: And just to be clear, you intend to return it Monday morning. **B:** Well, just to be on the safe I think I'll keep it for the week and return it Wednesday.
8	A: Now I'll need to take some details from you so that I can fill out all the paperwork. Your name is? **B:** Monica Beebor. A: Okay. And could you give me your address please? **B:** My old address or the new one? A: We'll take the old address then. We just need somewhere where we can send the confirmation of your booking to. **B:** My current postal address is 14 Castle Street. A: Next thing on the list is a telephone number. Given that you're in between houses. I think it makes sense to take your cell phone number. We need a means to contact you in case of emergency. **B:** That's 0231-4463-7689.

9	A: Now I need some details about your driving record for insurance purposes. I take it you have a full driving licence and not a provisional one. **B:** Yup that's right. You don't need a special licence for that do you? A: No. no. Anyone with a full driving licence is qualified to drive a van of that size. Any convictions? **B:** Nothing serious. Just a few parking tickets A: Okay. They don't count. So that is no convictions. Have you had any insurance claims in the past 5 years? **B:** Just the one when a car hit me while I was stationary in a car park.
10	A: Almost there now. You'll need to show some ID when you collect the van. Anything with your photo on it will do just fine. **B:** I carry my passport with me all the time so I'll use that. A: And how will you be paying Ma'am? You don't have an account with us, do you? We take all the major credit cards. **B:** I'll use my Visa card. A: Can you give me your card number. **B:** The number is 2344-5587-6489-0112 A: Could you give me its expiry date please? **B:** It's December 2015. A: Thanks Ma'am. See you on Friday. **B:** Thanks. See you.

3- Business courses

11	A: Hello, this is Business Nationwide, Daniel speaking, how can I help you? **B:** Hi. I've recently started up a small business and I need some information about your courses. A: We offer two courses which may be of interest to you. Our first course is called 'Getting Started'. It's a two-hour evening course, and it runs from 6pm to 8pm. We discuss things like "Is starting a business right for me?", writing a business plan and some of the legal issues. It runs at various locations in the area. Where are you based? **B:** I live in Eastleigh. A: Eastleigh. So, the closest course to you would be in Handbridge, and the next one is on the 20th March. **B:** What are the topics of the course?
12	A: we discuss the following topics: "Creating a business case action plan", "Initiating a business case action plan", "Developing the initiative", "Analyzing your options", "Managing risk and Ranking alternative solutions". **B:** Sounds good, and how much is that. A: That one is free. Do you want to participate in that? **B:** Okay, What background do I need? A: This course is appropriate for anyone who wants to develop a clear business case for decision-making. There are no prerequisites. **B:** well it might be worth it.
13	A: But did you say you're trading already? When did you start your own business? **B:** Yes, since about August. A: Well, you might be better off taking our three-day course - 'Business Basics'. It's not free I'm afraid. It's subsided and costs $80 for the three days, unless you've been unemployed in the past six months, in which case it's just $20. **B:** No, that doesn't apply to me. A: Well, it's well worth the money. The three days cover the essential aspects of running a business. The first day covers legal issues, such as tax, insurance, employment laws and health and safety. The second day covers marketing and pricing, and the third covers accounting and book-keeping. **B:** It sounds useful. Does the âĂŸBusiness Basics' course take place in Handbridge too?

14	A: Let me see. No, it's not available in Handbridge, I'm afraid. The nearest course to you would be in Renton. There's one on the 5th March, and another on the 18th April. **B:** Yes, that might be useful. A: I'll send out a pack to you if you like, with some details of the courses and also some information about what you need to do to set up and who you need to register with. **B:** Great. A: Can I take your name? **B:** Yes, it's Lila Park.
15	A: And can you give me your address please? **B:** It's 39 White Lane, Eastleigh. A: And have you got an email address? If so, we can send you details of any courses that are happening near you that you might be interested in. **B:** Yes, it's lila dot park at rainbow dot com A: Great! well, I'll have the information pack sent out to you today. **B:** Thanks, that'd be great. A: My pleasure. Bye.

4- Adventure

16	A: Good morning, White Cloud Adventures. How can I help you? **B:** I need some information about excursions you offer in this region. A: Okay, well we offer a number of tours. Where were you thinking of going? **B:** the overnight trip that takes people to see dolphins A: Yes, that's our Dolphin Watch Explorer. It's very popular so you usually need to book in advance. **B:** How long does it last? A: It's actually 1.5 days and that includes the two hour travel time each way from here. **B:** we leave from this office and travel by minibus?
17	A: Yes, you leave from here, but you travel by coach, then split into smaller groups for the boat trip. There are up to 20 people on each boat, though we run tours for numbers over 15. The boats travel to different spots in the ocean, so we don't have too many people disturbing the dolphins. **B:** Okay so how regular are the tours? A: Well, because its peak time now there is a tour every day, though we run them once a month outside of peak season and at certain times of year we close the tours down completely for a while. **B:** when is the next available date?
18	A: Just let me check on the computer. Okay well the next available booking is six days from now on (Q3) December 18th and then we don't have any free places on another tour until December 23rd. **B:** is it a good time of year to see the dolphins? A: That's actually quite unusual, though of course the dolphins can be unpredictable! Actually this is a very good time of year to see them and we have a policy that if you don't get to see any throughout the 1 1/2 days you get a 50% discount on a second booking. **B:** Really? That's great!

19	A: Yes, and you get to stay in really nice accommodation too. I would say the majority of our customers rate the tour really highly. The hotel we use is called the Kaoriland Lodge, it's right on the beachfront and has a lovely restaurant and gardens. **B:** Sounds good. So what is actually included in the tour price? A: Almost everything! The price includes transportation to and from the hotel, the boat trip, overnight accommodation, lunch on the boat and breakfast at the hotel the next day. You do need to pay for your evening meal though, that isn't included in the price. The hotel restaurant is quite reasonably priced and there are one or two nice eating places just a few minutes walk away too, if you feel like exploring. **B:** What do I need to take with me for the boat trip?
20	A: Well drinks and lunch are included in the package, so you really just need to take your swimwear, a towel and some warm clothing as it can get quite windy on the boat. Don't forget your camera and some sun screen. Oh, and I forgot to mention you can hire snorkels and kayaks very cheaply while you're on the trip, so you can have fun in the water when the boat stops for lunch. **B:** Are there many things to do once the boat trip is finished? A: Well, you'll get back to the hotel at around 5.30pm, it has lovely hot pools and there is also a sauna and gym. You have to pay extra to use those. But there is also a floodlit tennis court and indoor swimming pool which are free to guests. **B:** So how much does the tour cost?
21	A: It is usually $420 at this time of year; but there is a discount for bookings of more than 3 people which brings the price down to $390 per person. **B:** So do you have enough spaces for four people on the next tour? A: Just let me check, I think that should be okay. ... Yes, that's fine. We have 6 places left at the moment. **B:** Okay well I'd like to book those places please. A: Alright. Well, you need to pay a 50% deposit at the time of booking. That is fully refundable if for any reason you have to cancel more than 48 hours before departure date. The full amount must be paid at least 2 days before departure. If you cancel within 48 hours of the tour, you have to pay 25% cancellation fee. **B:** can I pay by credit card?

22	A: Yes, credit card is fine. So I'll make a booking for four people for the 18th of December. You'll need to be here at this office by no later than 7.45am on the day of departure. The coach leaves promptly at 8am. **B:** Yes, okay, that's fine A: Well that's all booked for you now. Write down this booking number too, just in case you lose your tickets or anything. **B:** Okay, what is it? A: It's 2-6-5-9-1-T. That's T for tango. That's all done for you. Hope you have a great trip! **B:** Thanks very much, I'm sure we will.

5- Book a hotel

23	A: Worldbridges Travel Agency. Good morning. Can I help you? **B:** We are three friends..., and we'd like to travel to Greece next July A: Quite difficult in July, would you like to stay at a hotel? Or are you thinking of a villa or an apartment? **B:** I guess a small apartment will be cheaper A: Provided it's not during July, yes. You know prices are lower out of season. How long would you like to stay? **B:** About a fortnight and we cannot spend more than $100 a day...
24	A: For that price you won't have many options, I'm afraid, but let me find out. If you could arrange to make it in late June. I might have a bedsitter for $75. It could accommodate 3 single beds, and it's 5 minutes' walk from the main beach in Mykonos. **B:** I'd love that. I need to talk to my friends. A: Well, contact your friends, come to an agreement and give me a ring again. My name is Arnold Smith, you'll find me here any working day from 10 am to 6 pm, but not on Saturdays. Remember we only have a month left, so you need to make up your minds I'd say today or tomorrow! **B:** I will, thank you A: Wait ! You haven't given me your name. **B:** Sorry..., I am Susan Perkins A: Thanks. **B:**Thank you again and bye.

6- Book a table in a restaurant

25	A: Hello, Fusion Restaurant. How can I help you? **B:** Hello, do you do group bookings? A: Yes, we do. **B:** Well, I'm not exactly sure of numbers right now but I'd like to book a table for between 15and 20 of us. A: When is it for? **B:** A week on Friday. A: The 16th of November? **B:** Yes, that's right. A: OK, well, usually we offer a set menu for groups up to 25 people. If you want, you can order a la carte, but it usually works out more expensive and obviously if everybody orders different food, then it can take us a lot longer to prepare. And particularly on a Friday as we're usually pretty busy. **B:** So, would you recommend the set menu?
26	A: Obviously it's up to you, but it's a good option and you get to try different types of food. **B:** That's good. We have a couple of vegetarians in the group. another who is allergic to peanuts A: That's not a problem. We cater for all dietary requirements. We offer several different menus and you can choose which one you prefer. Our basic menu costs $25 per person. We do it for a minimum of 4 people, as with all our menus. That's menu A. Then they each go up in price depending on which one you order. Menu B is $30 per person. Menu G is $35 per person and menu J is $40 per person. **B:** And how do the menus differ?
27	A: Well, all our set menus include a variety of food âĂŞ some Asian style, some Mediterranean, some Latin American and some British food. **B:** And can you mix all types of food in each menu? A: Yes, that's right. Basically you get a selection of starters for the whole table, for example in menu A, you get a goat's cheese salad, onion bhajis and guacamole. With menu J, you get spare ribs, king prawns, hummus and pitta bread and a selection of salads. You get more options with the more expensive menus. **B:** What about the main courses?

28	A: You have a choice of 3 options. All menus also include a dessert and coffee. **B:** Right, I see. Could I pop by and pick up some menus to have a look at? A: Of course. Or I could email them to you, if you want, to save you the trouble. **B:** That'd be great. A: So, what's your name? **B:** My name is Hannah Bailey A: OK. And your e-mail? That's hb0470@freemail.com. And my phone number is 01793 211873. A: All right, I will make a temporary reservation for you? how many people did you say? **B:** Well, I'm not sure, between 15 and 20. A: OK, that's fine. You can confirm numbers and which menus you want a couple of days before and I'll send you the email with the menus right now. **B:** OK, thanks a lot.

7- Confirming a hotel booking

29	A: Good afternoon, Orion Hotel, how may I help you? **B:** I'm ringing to confirm a booking I made a week ago. I was expecting an email but I haven't received anything. A: I'm sorry the line is rather bad, would you mind repeating that, please? **B:** Yes, I made a reservation on your website under the name of Coutts. Sandra Coutts. A: And when was the reservation for? **B:** July 23^{rd} to the 29^{th} A: Let me just check if we have your details on the system. Yes, here we are. **B:** I put down one double and one single room, but I wonder if I could change that. A: I'm just getting your booking details up now. Yes, we do have another double available for those nights. Was there anything else? **B:** Well, I wanted to know if you did vegetarian food for the evening meal. It wasn't clear from the website.
30	A: Yes, we do, but by prior arrangement. But now you've requested it, I'll put that down in the booking. How many people was that for? **B:** Just myself. A: Very good. Anything else? **B:** Well, we've got a very early return flight on the 30th, so we'll need an alarm call at about 5.30, I should think. A: No problem sir, So you'd like two double rooms for the nights of 23rd to the 29th July inclusive, vegetarian provision for one and an early morning call on your departure. Is that correct? **B:** That's right A: If we can help you with anything else, just give us a ring. We look forward to seeing you in July. **B:** Thanks. Bye.

CURRICULUM VITA
NZANIN POURMAND

EDUCATION

2008-2012 — **The University of Western Ontario**
Degree: **Ph.D. in Electrical Engineering** — **London, Canada**

2004-2007 — **Amirkabir University of Technology**
Degree: **M.Sc. in Electrical Engineering** — **Tehran, Iran**

1999-2003 — **Isfahan university of Technology**
Degree: **B.Sc. in Electrical Engineering** — **Isfahan, Iran**

THESES

- Ph.D. Thesis
  Objective and Subjective Evaluation of Wideband Speech Quality
  Advisor: Dr. Vijay Parsa

- M.Sc. Thesis
  Improvement of Soft Stochastic Segment Modelling Performance in Persian Acoustic to Phonetic Decoder for Continuous Speech
  Advisor: Dr. A. Sayadiyan

- B.Sc. Thesis
  Speaker Identification Using Wavelet Transform
  Advisor: Dr. S. Sadri

TEACHING EXPERIENCE

2007-2011 — **The University of Western Ontario**
Position: **Teaching Assistant** — **London, Canada**

WORK EXPERIENCE

2008–2012 — **National Centre for Audiology**
Position: **Research Assistant** — **London, Canada**

Summer 2011 — **Research in Motion Inc.**
Position: **Research intern** — **Waterloo, Canada**

Summer 2004 — **Foolad Technic Co.**
Position: **R&D Engineer** — **Isfahan, Iran**

Summer 2003 — **SarvNet Telecommunications Inc.**
Position: **Research Intern** — **Isfahan, Iran**

HONORS AND AWARDS

- Western Graduate Research Scholarship (WGRS), UWO, London, Canada [2008-2012].
- Best presentation award in graduate symposium, Electrical and Computer Engineering Department, UWO [2010].
- Graduated with distinction from Amirkabir University of Technology [2007].

JOURNAL PAPERS

- **N. Pourmand**, and V. Parsa, A. Weaver "Computational auditory models in predicting noise reduction performance for wideband telephony applications", *International Journal of Speech Technology*, 2013, 1-17.

REFEREED CONFERENCE PAPERS

- **N. Pourmand**, D. Suelzle, V. Parsa, Y. Hu, P. Loizou,"On the use of Bayesian modeling for predicting noise reduction performance", *IEEE International Conference on Acoustics, Apr. 2009*.
- **N. Pourmand**, A. Sayadiyan,"Evaluation of soft segment modeling in vowel detection and classification in Persian discrete speech", *Iranian Conference on Electrical Engineering, Tehran, Iran, May 2007*.
- **N. Pourmand**, A. Sayadiyan,"Vowel detection and classification in Persian discrete speech with combination of acoustic method and soft segment modeling", *International CSI Computer Conference, Shahid Beheshti University, Tehran, Iran, Feb. 2007*.

PEER REVIEWED ABSTRACTS

- **N. Pourmand**, V. Parsa, A. Cowley, M. Gupta, and C. Forrester, "Objective and subjective speech quality evaluation of wideband noise reduction algorithms", *Acoustical Society of America meeting, Cancun, Mexico, Nov. 2010*.
- **N. Pourmand**, V. Parsa, "Objective assessment of speech enhancement algorithms", *Canadian Acoustical Association Conference, Niagara-on-the-lake, Canada, Oct. 2009*.

BOOK

- **N. Pourmand** *et al.*, "The complete solutions to probability problems", Arkan Publications, Isfahan, Iran, 2005, ISBN: 964-7983-25-5.