Blind Estimation of Perceptual Quality for Modern Speech Communications

by

Tiago Henrique Falk

A thesis submitted to the
Department of Electrical and Computer Engineering
in conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
December 2008

Abstract

Modern speech communication technologies expose users to perceptual quality degradations that were not experienced earlier with conventional telephone systems. Since perceived speech quality is a major contributor to the end user's perception of quality of service, speech quality estimation has become an important research field. In this dissertation, perceptual quality estimators are proposed for several emerging speech communication applications, in particular for *i*) wireless communications with noise suppression capabilities, *ii*) wireless-VoIP communications, *iii*) far-field hands-free speech communications, and *iv*) text-to-speech systems.

First, a general-purpose speech quality estimator is proposed based on statistical models of normative speech behaviour and on innovative techniques to detect multiple signal distortions. The estimators do not depend on a clean reference signal hence are termed "blind." Quality meters are then distributed along the network chain to allow for both quality degradations *and* quality enhancements to be handled. In order to improve estimation performance for wireless communications, statistical models of noise-suppressed speech are also incorporated.

Next, a hybrid signal-and-link-parametric quality estimation paradigm is proposed for emerging wireless-VoIP communications. The algorithm uses VoIP connection parameters to estimate a base quality representative of the packet switching network.

Signal-based distortions are then detected and quantified in order to adjust the base quality accordingly. The proposed hybrid methodology is shown to overcome the limitations of existing pure signal-based and pure link parametric algorithms.

Temporal dynamics information is then investigated for quality diagnosis for hands-free speech communications. A spectro-temporal signal representation, where speech and reverberation tail components are shown to be separable, is used for blind characterization of room acoustics. In particular, estimators of reverberation time, direct-to-reverberation energy ratio, and reverberant speech quality are developed.

Lastly, perceptual quality estimation for text-to-speech systems is addressed. Text- and speaker-independent hidden Markov models, trained on naturally produced speech, are used to capture normative spectral-temporal information. Deviations from the models, computed by means of a log-likelihood measure, are shown to be reliable indicators of multiple quality attributes including *naturalness*, *fluency*, and *intelligibility*.

To my beloved wife, whose beautiful smile has the power to transform even the darkest days into sunshine.

Acknowledgements

First, I would like to express my sincerest gratitude to Dr. Geoffrey Chan, whose profound insights and knowledge have greatly influenced my academic career and have helped me become an independent researcher. I still recall our first meeting when Dr. Chan mentioned that he would "throw me in the middle of the ocean" and my mission as a Doctoral student was to swim back to shore with as little assistance as possible. It feels good to be writing this Acknowledgements section, as it signals my safe arrival back to shore!

I would also like to thank Dr. Douglas O'Shaughnessy, Dr. Ingrid Johnsrude, Dr. Fady Alajaji, and Dr. Michael Greenspan for taking time off their busy schedules to serve on my Thesis Examination Committee and also for their valuable comments. Sincere thanks are also expressed to Dr. Bastiaan Kleijn, Dr. Sebastian Möller, and Dr. Peter Kabal for fruitful collaborations, discussions, and suggestions that helped improve this dissertation.

This thesis would not have been concluded if it were not for the numerous colleagues who kindly assisted with data collection and/or generously donated speech databases. In particular, I would like to express my gratitude to Dr. Leigh Thorpe, Mr. Jimi Wen, Mr. David Gelbart, Dr. Koen Eneman, Dr. Yi Hu, Dr. Philip Loizou, Dr. Roch Lefebvre, Mr. Stephen Voran, Dr. Doh-Suk Kim, Dr. Samy El-Hennaway, and Dr. Lijing Ding. I would also like to thank my many friends – too numerous to list here – who, aside from assisting with research, also helped keep me sane during these last four years. You know who you are; Thanks!

Graduate studies would not have been fun if it were not for the wonderful and cheerful ECE staff. Debie, Bernice, Tom, Kendra, and Patty, I will greatly miss our conversations and get-togethers. Bernice, I wish you all the best on your retirement. I would also like to acknowledge the funding received from the Natural Sciences and Engineering Research Council of Canada (NSERC) throughout my studies.

I would not have been here today if it were not for the love and care of my family. I would like to greatly acknowledge my parents who have always supported, encouraged and taught me to be the person I am today. I am also grateful to my siblings who have always stood by my side. To my wife – my better half, my soul mate – no words can express how grateful I am for your endless love, support, and encouragement. To you, and to the tiny one the world has yet to meet, I dedicate this thesis.

"When one has many people to thank, one must first thank God."

Anonymous

Statement of Originality

I hereby certify that all of the work described within this dissertation is the original work of the author. Any published or unpublished ideas and/or techniques from the work of others are acknowledged in the references described herein.

Tiago H. Falk

Kingston, ON

Contents

List of Tables

List of Figures

List of Abbreviations

ACR	Absolute Category Rating
AMR	Adaptive Multi Rate
ANIQUE+	Auditory Non-Intrusive Quality Estimation Plus
ANSI	American National Standards Institute
BSD	Bark spectral distortion
CART	Classification and Regression Tree
CCR	Comparison Category Rating
CQE	Conversational Quality Estimate
CQO	Conversational Quality Objective
CQS	Conversational Quality Subjective
DAM	Diagnostic Acceptability Measure
DCR	Degradation Category Rating
EM	Expectation-Maximization
DRR	Direct-to-reverberation energy ratio
ETSI	European Telecommunications Standards Institute
EVRC	Enhanced Variable Rate Codec
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IP	Internet Protocol
IR	Impulse Response
ITU-T	International Telecommunications Union – Telecommunications
KLD	Kullback-Leibler Distance
LPC	Linear Prediction Coefficient
LQE	Listening Quality Estimate
LQO	Listening Quality Objective
LQS	Listening Quality Subjective
MAE	Median Absolute Error
MARDY	Multi-channel Acoustic Reverberation Database at York

MARS	Multivariate Adaptive Regression Spline
MFCC	Mel-frequency Cepstral Coefficient
MNB	Measuring Normalizing Block
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
MSE	Mean-square Error
ORSMR	Overall Reverberation-to-Speech Modulation energy Ratio
PC	Personal Computer
PESQ	Perceptual Evaluation of Speech Quality
PLC	Packet Loss Concealment
PLP	Perceptual Linear Prediction
POTS	Plain Old Telephone System
PSQM	Perceptual Speech Quality Measure
QoS	Quality of Service
RBF	Radial Basis Function
RFC	Random Forest Classifier
RMSE	Root-mean-square Error
RSMR	Reverberation-to-Speech Modulation Energy Ratio
RTP	Real-time Transport Protocol
RTCP	Real-time Transport Control Protocol
RTCP-XR	Real-time Transport Control Protocol – Extended Report
SRMR	Speech-to-Reverberation Modulation energy Ratio
SMV	Selectable Mode Vocoder
SNR	Signal-to-Noise Ratio
SVC	Support Vector Classifier
SVR	Support Vector Regressor
TTS	Text-to-Speech
VAD	Voice Activity Detection
VoIP	Voice Over Internet Protocol

Chapter 1

Introduction

The speech communications industry is going through a phase of rapid development and new services and technologies are emerging continuously. In a society where "mobility," "low-cost," and the ability to "multi-task" have become essential, a decline in the use of the plain old telephone system (POTS) has been witnessed. As an example, a recent study by Statistics Canada has shown that "cellphone use in Canada has just about caught up with traditional wire lines as the wireless industry continues to grow in reach and profitability" [1]. The report shows that in December 1999, of 100 Canadian inhabitants, 18.7 were wireless subscribers and 64.4 were traditional wireline subscribers. As of December 2006, these percentages were 55.1 and 55.3, respectively.

Where low-cost telephony is concerned, voice over internet protocol (VoIP) has been gaining grounds rapidly. Recent technologies include cable VoIP [2], mobile VoIP (also known as wireless-VoIP) [3–5], as well as conventional VoIP, where service providers, such as Skype and Vodafone, have gained wide popularity. According to recent studies, users can expect savings of up to 40% on telephone bills by switching to VoIP; businesses can expect larger savings with substantially lower long-distance expenses [6]. Moreover, companies such as AwayPhone claim that savings of up to 90% in mobile phone calls made from abroad can be attained with the use of Mobile VoIP [7]. In fact, recent research by the consulting firm ON World has suggested that by 2011 the number of wireless-VoIP users around the world will rise to 100 million from 7 million in 2007. It is also estimated that in 2011, wireless-VoIP voice services will generate $33.7 billion, up from $516 million in 2006, the most recent year for which the figure is available [8].

Although mobility and low-cost seem to be the driving forces behind the expansion of wireless and VoIP services, multi-tasking (in conjunction with mobility) is paving the way for hands-free speech communications. Applications include voice-activated controls in automobiles, voice controlled applications in personal computers (PC), as well as conventional applications such as video conferencing. More recently, popular VoIP PC-based telephony applications such as Skype, MSN Messenger, and GoogleTalk have also increased demand for hands-free communications. In such applications, the use of a single microphone is not efficient and, commonly, multiple microphones (also known as microphone arrays) are used to reduce background noise and reverberation. As examples of the expected growth and popularity of hands-free communications, Microsoft's new operating system, Windows Vista, provides support for microphone arrays; moreover, most computer manufacturers are now producing laptops that are equipped with a microphone array.

While such technological advances facilitate human interaction, users are now exposed to perceptual degradations that were not experienced with the conventional POTS. Examples of such distortions include varying types and levels of acoustic background noise, packet losses, and reverberation, all of which have peculiar characteristics that are detrimental to speech quality and intelligibility. In order to reduce such detrimental effects, current research has focused on developing novel "speech enhancement" algorithms for acoustic noise suppression, packet loss concealment (PLC), and reverberation suppression (also termed *dereverberation*). Under severe adverse conditions, however, even the most advanced state-of-the-art speech enhancement algorithm will undoubtedly introduce unwanted perceptual artifacts that compromise speech quality and intelligibility.

Moreover, a research area that has also witnessed rapid growth over the last decade is that of text-to-speech (TTS) synthesis. As the name suggests, TTS systems attempt to convert arbitrary input text into intelligible and naturally sounding speech. Earlier applications of TTS systems served mostly as an aid to the visually impaired. Today, TTS systems have broad applications in education, business, entertainment, and medicine. Representative applications include email and short message service readers, automated directory assistance, foreign language education [9], and assistive and augmentative communications [10]. As reported in recent Blizzard TTS Challenges,[1] current state-of-the-art TTS systems, albeit producing high-quality naturally sounding outputs, are still not capable of synthesizing speech that is indistinguishable from naturally-produced speech.

Ultimately, the success or failure of an innovative speech communication technology relies on the end user's perception of "quality" and "usability." While the latter can comprise factors such as cost and ease-of-use, the former commonly includes factors such as presence of perceptual artifacts, (un)naturalness of the speech signal, or

[1]The Blizzard Challenge is run annually by the ISCA Speech Synthesis Special Interest Group and consists of a venue where different TTS systems are subjectively evaluated and compared.

loss in speech intelligibility. In this manuscript style thesis, focus is placed on *quality* evaluation methods for *emerging* speech communication applications. Quality evaluation can be performed either subjectively, with human listeners, or objectively, by means of a computational algorithm. In Sections 1.1-1.2 to follow, subjective and objective speech quality assessment methods are reviewed. Motivation and objectives, thesis contributions, and thesis organization are further described in Sections 1.3-1.5, respectively.

1.1 Subjective Speech Quality Assessment

As defined in [11], speech quality is the result of a subjective perception-and-judgment process, during which a listener compares the perceptual event (speech signal heard) to an internal reference of what is judged to be "good quality." Subjective assessment plays a key role in characterizing the quality of emerging telecommunications products and services. It is known, for example, that the perceived quality of a speech signal processed by a novel speech coding algorithm, or transmitted over a novel network architecture, will reflect the end user's experience with the system under test. Subjective speech quality testing attempts to quantify this user experience. Moreover, the results of subjective evaluations can be used to define performance targets, to ensure appropriate product performance, and to define national and international standards [12].

Subjective tests can be grouped into two larger classes: listening- and conversational-quality tests. Listening tests, as the name suggests, has listeners "passively" rate (on a pre-specified scale) the quality of the short-duration speech signal they have just heard. Conversational tests, on the other hand, are interactive and listeners are asked to rate the quality of a call based on the listening quality *and* on their ability to converse during the call. In conversational tests, factors such as echoes and delays have to be taken into account. Listening quality tests are by far the most widely used tests in the speech communications realm and can be further classified as unidimensional or multidimensional. The International Telecommunication Union (ITU-T) has, over the years, published several Recommendations describing guidelines for conducting uni- and multidimensional subjective evaluations of listening quality in order to obtain reliable and reproducible test results (e.g., see [13–15]). Some representative subjective listening quality tests are described next.

1.1.1 Unidimensional Listening Quality Tests

ITU-T Recommendation P.800 [13] describes three unidimensional scales to be used for subjective *listening* quality tests: absolute category rating (ACR), degradation category rating (DCR) and comparison category rating (CCR). In ACR testing, listeners are instructed to rate the processed speech material presented to them according to the 5-point quality scale described in Table 1.1, column labeled "ACR." Listeners are not presented with clean reference speech files for comparisons and are asked to rate the "absolute" quality of the speech samples. The average of the listener scores is termed the subjective mean opinion score, or subjective MOS.

On the other hand, in DCR and CCR tests listeners are presented with both the reference (clean) and the processed (degraded) speech signals. With DCR tests, listeners are instructed to rate the perceived degradation of the processed speech material relative to the unprocessed material using the scale shown in Table 1.1, column labeled "DCR." In the CCR test, listeners are asked to identify the quality of

Table 1.1: Subjective rating scale for absolute category rating (ACR) and degradation category rating (DCR) tests.

Rating	ACR	DCR (Level of degradation)
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

the processed speech sample relative to its unprocessed counterpart using a two-sided rating scale, as given by Table 1.2. During half of the trials, the unprocessed sample is followed by the processed sample; the order is reversed for the remaining trials. As such, CCR testing improves on DCR testing as it minimizes biases that occur due to the order in which the speech materials are presented.

Since DCR and CCR tests require the presentation of two speech signals per trial, they usually take longer to perform. According to [12], listeners are able to attend to speech samples and give consistent ratings for tests up to about an hour long. As a consequence, subjective listening tests usually span multiple sessions over multiple weeks. This limitation, in combination with the costs associated with multiple-session tests, has popularized ACR testing. Today, it is the most common type of listening test in the telecommunications industry; the abbreviation MOS-LQS is commonly used to denote listening quality subjective MOS [15].

1.1.2 Multidimensional Listening Quality Tests

A major drawback of the aforementioned subjective tests is that listeners rate the quality of the speech signal using a single perceptual quality dimension. Studies have

Table 1.2: Subjective rating scale for comparison category rating (CCR) tests. Listeners are asked to rate the quality of the signal played the second time relative to the signal played the first time.

Category	CCR
3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

shown that different dimensions are involved in subjective quality perception [16–18]. Commonly, multidimensional tests solicit separate reactions from the listeners with regards to what is perceived as the speech signal itself, the background, and the overall effect [16]. The so-called diagnostic acceptability measure (DAM) test, proposed in the late 1970's, was the first test to categorize distortions according to several different attributes [16].

The DAM methodology evaluates the signal using thirteen separate diagnostic scores: six are based on perceptual qualities of the *signal*, four on perceptual qualities of the *background*, and three on perceptual qualities of the signal-plus-background *total effect*. All diagnostic scores are rated on a monopolar rating scale ranging from 0 (negligible effect) to 100 (extreme effect). A description of the different attributes that are evaluated in a DAM test is given in Table 1.3, along with each attribute's intrinsic effect on acceptability. Composite *signal* and *background* perceptual quality scores can be obtained by aggregating individual scores obtained by the attributes in each domain. Lastly, a single-dimensional (overall) acceptability measure may be obtained by aggregating the composite scores and the *total effect* scores.

Table 1.3: Diagnostic acceptability measure quality scales

Perceptual domain	Descriptor	Effect
Signal	fluttering, bubbling	severe
	distant, thin	mild
	rasping, crackling	severe
	muffled, smothered	mild
	irregular, interrupted	moderate
	nasal, whining	moderate
Background	hissing, rushing	moderate
	buzzing, humming	moderate
	chirping, bubbling	severe
	rumbling, thumping	moderate
Total effect	intelligibility	–
	pleasantness	–
	acceptability	–

Due to time and cost constraints, multidimensional tests did not gain popularity and were seldom used by the communications industry in the last 30 years. Advances in speech communications and speech enhancement technologies, however, have revived interest in multidimensional tests. As an example, the study described in [18] proposes to categorize modern distortions according to thirteen different labels (e.g., noisy, muffled, intelligible, interrupted). Via principal components analysis, the thirteen descriptors are shown to group into three quality dimensions, namely: "directness/frequency content," "continuity," and "noisiness." Regression analysis has shown that continuity appears to be the most important dimension in terms of overall listening quality [18].

Moreover, the ITU-T has recently published Recommendation P.835 [19], which sets guidelines as to how to conduct multidimensional tests for systems that include an acoustic noise suppression algorithm. It is known, for example, that many noise

Table 1.4: Subjective rating scale for signal distortion (SIG) and background intrusiveness (BCK), according to ITU-T Rec. P.835

Rating	SIG	BCK
5	Not Distorted	Not Noticeable
4	Slightly Distorted	Slightly Noticeable
3	Somewhat Distorted	Noticeable but Not Intrusive
2	Fairly Distorted	Somewhat Intrusive
1	Very Distorted	Very Intrusive

suppression algorithms can introduce unwanted artifacts to the speech signal. A typical artifact is known as "musical noise" [20]. In such situations, subjects can become confused as to which components of a noisy speech signal should form the basis of their ratings of overall quality. To reduce the error variance (or listener uncertainty) in the subjects' ratings of overall quality, the P.835 methodology instructs the listener to successively attend to and rate three different components of the noise suppressed speech signal: the speech signal alone, background noise alone, and the overall effect.

When assessing the speech signal alone, listeners are instructed to use the "signal distortion" scale described by column "SIG" in Table 1.4. The background noise is examined using the "background intrusiveness" scale described by column "BCK." The overall effect uses the ACR scale described in Table 1.1 and the notation "OVRL" will be used throughout the remainder of this dissertation. The process of rating the signal alone and the background noise alone leads the listener to integrate the effects of both the signal and the background in making their ratings of overall quality [21].

Although multidimensional tests have not been formally proposed to specifically handle distortions resulting from packet losses (and packet loss concealment algorithms) or reverberation (and dereverberation algorithms), the subjective test described in [22], tuned to reverberant and dereverberated speech, is a step forward in this direction. In the study, listeners are instructed to rate the quality of the signal based on three dimensions: colouration (frequency distortion due to early reflections, causes signal to sound "boxy"), reverberation tail effect (temporal smearing due to late reflections, causes signal to sound distant and with echoes), and overall speech quality. The methodology and guidelines described in [19] were used to carry out the subjective tests.

For synthesized speech, on the other hand, multidimensional quality tests, such as those described in ITU-T Recommendation P.85 [23], have been used since the early 1990's. In the test, listeners are asked to rate the signal using eight quality dimensions labeled: *overall impression* (MOS), *listening effort* (LSE), *comprehension problems* (CMP), *articulation* (ART), *pronunciation* (PRO), *speaking rate* (SRA), *voice pleasantness* (VPL), and *acceptance* (ACC). The "overall impression" rating uses the ACR scale shown in Table 1.1. The "acceptance" dimension, in turn, uses a two-point scale (yes/no) and results are reported as a percentage acceptance value. The scales used for LSE and SRA are reported in Table 1.5 and the remaining four quality dimensions are reported in Table 1.6.

During the test, subjects are presented with each synthesized speech file twice. In the first presentation, subjects are asked to solve a secondary task such as answer specific questions about information contained in the file (e.g., bus number and bus date/time of departure). Subjects are then asked to judge the quality of the speech signal based on the aforementioned quality dimensions. The intent of providing a secondary task is to direct the listeners' attention to the content of the speech signal, and not on its surface form alone, so as to improve listener judgement of e.g., comprehension problems and listening effort.

Table 1.5: LSE and SRA rating scales, as described in ITU-T P.85.

Rating	LSE [a]	SRA [b]
5	Complete relaxation possible; no effort required	Much faster than preferred
4	Attention necessary; no appreciable effort required	Faster than preferred
3	Moderate effort required	Preferred
2	Effort required	Slower than preferred
1	No meaning understood with any feasible effort	Much slower than preferred

[a]Listeners are asked to describe the effort required to understand the message
[b]Listeners are asked to rate the average speed of delivery

Table 1.6: CMP, ART, PRO, and VPL rating scales, as described in ITU-T P.85.

Rating	CMP [a]	ART [b]	PRO [c]	VPL [d]
5	Never	Yes, very clear	No	Very pleasant
4	Rarely	Yes, clear enough	Yes, but not annoying	Pleasant
3	Occasionally	Fairly clear	Yes, slightly annoying	Fair
2	Often	No, not very clear	Yes, annoying	Unpleasant
1	All of the time	No, not at all	Yes, very annoying	Very unpleasant

[a]Listeners are asked to rate if certain words were hard to understand
[b]Listeners are asked if the sounds were distinguishable
[c]Listeners are asked if they noticed any anomalies in pronunciation
[d]Listeners are asked how they would described the pleasantness of the voice

1.1.3 Conversational Quality Tests

With conversational quality testing, the listeners rate the quality of a call based on the listening quality *and* their ability to converse during the call. Conversational tests take into account echoes and delays as possible degradations of conversational quality. In conversational tests, listeners are placed into interactive communication scenarios and asked to complete a task over the phone. An example includes a proof-reading task in which each subject has a slightly different version of the same text and they are asked to find the differences [24]. By setting a time limit, the performance of the system can be measured indirectly based on the outcome of the task.

Moreover, at the end of each conversation, listeners are presented with questions regarding the quality of the system. One question focuses on the quality of the connection and the ACR scale shown in Table 1.1 is used. A second question addresses the difficulty in talking or hearing over the connection; a binary (yes or no) "difficulty" opinion scale is used in this case. In this scenario, the percentage "difficulty," or percentage of listeners who answered that they had difficulty in hearing or talking, is calculated over all listeners. To avoid confusion, ITU-T Recommendation P.800.1 [15] has introduced the abbreviation "MOS-CQS" to distinguish the mean opinion scores obtained with subjective conversational quality tests from those obtained via subjective listening quality (MOS-LQS) tests.

As can be seen, subjective tests have to be conducted following strict guidelines; such requirements are necessary in order to obtain accurate and repeatable results. Moreover, in order to reduce the effects of interlistener variability, subjective tests typically involve over 32 naïve listeners. Studies show that with such a sizeable listener panel, the 95% confidence interval of a MOS-LQS test (for a given degradation condition) is approximately 0.1 MOS [25]. In practice, this seldom represents a problem since even expert listeners would struggle to distinguish differences in quality of this magnitude in the ACR context [25]. Unfortunately, these requirements make subjective tests very expensive and time consuming, thus unsuitable for online applications or for frequent system evaluation, as occurs during the development and fine-tuning of TTS systems. Today, most of the research on speech quality measurement focuses on identifying and modeling audible distortions through an objective process. Objective methods can be implemented by computer programs and can be used in real-time measurement of speech quality. Objective speech quality measurement is the main topic of this thesis and special emphasis is placed on objective measures of *listening* quality. An overview of existing objective measures is described next.

1.2 Objective Speech Quality Assessment

Objective machine-based quality measurement allows computer programs to automate speech quality measurement in real-time, making it suitable for field and/or frequent applications. In fact, objective measurement is the only viable means of measuring voice quality, for the purpose of real-time call monitoring, on a network-wide scale. Machine-based algorithms aim to deliver estimated quality scores that are highly correlated with the quality scores obtained from subjective listening experiments. Objective measurement methods can be classified as either signal-based, parameter-based, or hybrid signal-and-link-parametric.

Signal-based methods (Fig. 1.1 (a,b)) use perceptual features computed from the speech signal to estimate subjective quality. Parameter-based methods (Fig. 1.1 (c)),

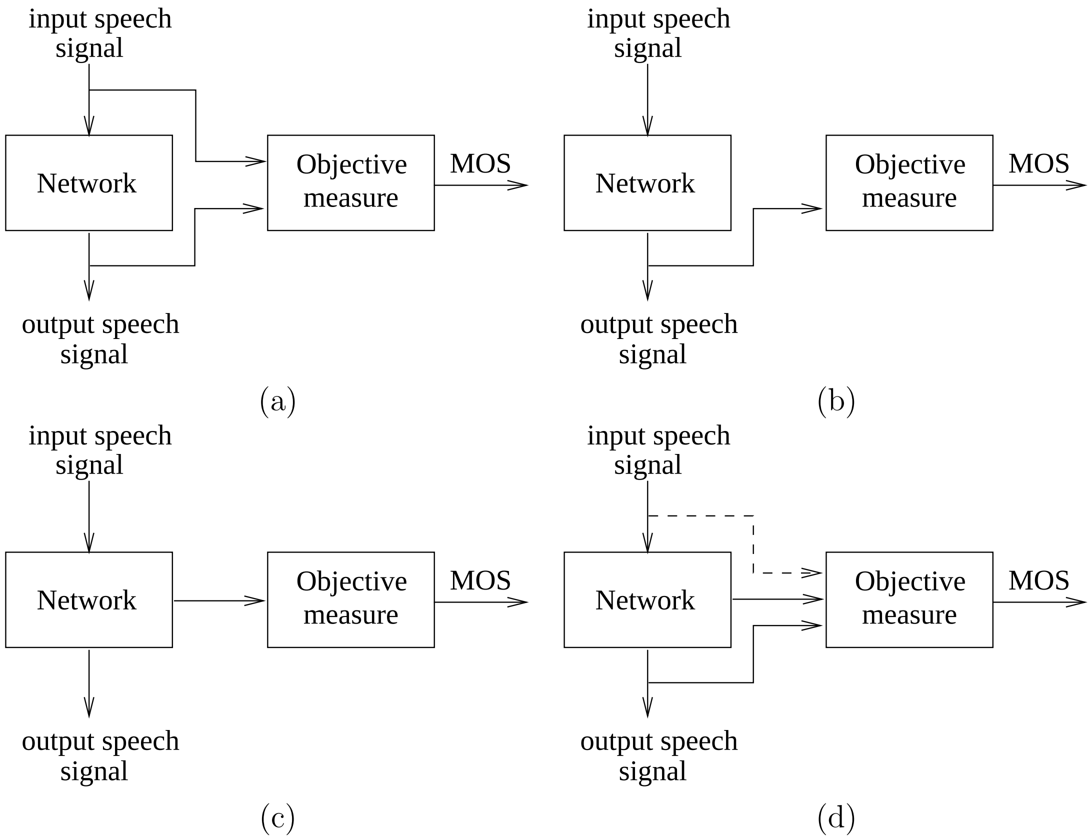


Figure 1.1: Block diagram of (a) double-ended and (b) single-ended signal-based objective measurement, (c) parameter-based measurement, and (d) hybrid signal-and-link-parametric measurement.

on the other hand, use either system or network parameters to estimate quality. Hybrid methodologies (Fig. 1.1 (d)) use both the signal payload and system and/or network parameters. Current MOS terminology recommends the use of the abbreviation MOS-LQO for “objective” listening quality MOS obtained from signal-based or hybrid models and the abbreviation MOS-LQE for “estimated” planning MOS obtained by parameter-based models [15]. Similarly, abbreviations MOS-CQO and MOS-CQE are used for conversational quality.

1.2.1 Signal-Based Measures

Signal-based quality measurement methods can be further classified as double-ended or single-ended based on the input information that is required. Double- and single-ended measurement paradigms are depicted in Fig. 1.1 (a) and (b), respectively. Double-ended measurement systems are "comparison-based" and depend on some form of distance metric between two input signals – a reference (clean) and a degraded speech signal at the output of the system under test – to estimate MOS-LQS. Single-ended measurement, on the other hand, depends only on the degraded speech signal and constitutes a more challenging paradigm. Single-ended measures, also commonly referred to as "blind," "reference-free," or "single-input" estimators, are the focus of this dissertation. These terms will be used interchangeably throughout the remainder of this thesis. An overview of double- and single-ended measurement methods is presented next.

1.2.1.1 Double-Ended Algorithms

Double-ended quality measurement has been studied since the early 1980's [26]. Earlier methods were implemented to assess the quality of waveform-preserving speech coders; representative measures include signal-to-noise ratio (SNR) and segmental SNR [27]. More sophisticated measures (e.g., [28]) were proposed once low bitrate speech coders, which may not preserve the original signal waveform, were introduced. More recently, quality measurement research has focused on algorithms that exploit models of human auditory perception. Representative algorithms include Bark spectral distortion (BSD) [29], perceptual speech quality measure (PSQM) [30], measuring normalizing block (MNB) [31, 32], and statistical model-based quality measurement

[33–35]. The International Telecommunications Union ITU-T P.862 standard, also known as perceptual evaluation of speech quality (PESQ), represents the current state-of-the-art double-ended algorithm [36]. Recent research, however, has suggested decreased PESQ performance for VoIP communications and algorithm sensitivity to connection parameters such as speech codec and PLC type, packet size, packet loss rate, and packet loss pattern (see e.g., [37–39]).

Moreover, double-ended schemes have two underlying requirements: (1) that the input (reference) signal be of high quality, i.e., clean, and (2) that the output (processed) signal be of quality no better than the input. These requirements prohibit the use of double-ended algorithms in scenarios where the input is degraded and the system being tested is equipped with a speech enhancement algorithm. Hence, the use of double-ended measures for modern speech communications is questionable, as will be emphasized in Chapter 3. In fact, ITU-T Recommendation P.862.3 [40] states that "the use of PESQ with systems that include noise suppression algorithms is not recommended."

The need for a reference signal also compromises the usability of double-ended systems for quality measurement of synthesized speech, as a "clean" reference signal may only be existent with corpus-based concatenative systems. With such systems, signal-based measures have been proposed and focus on computing spectral distances between the target synthesized speech signal and its original natural speech counterpart (see e.g., [41, 42]). Such measures, however, are only useful if perceptual degradations are linked to concatenation effects and if a reference natural speech corpus is available; such requirements are not always met in practice [43]. Alternately, the work described in [44] proposes the use of PESQ for quality estimation. In the study, the natural speech signal uttered by the same speaker with which the TTS corpus was built from is used as a reference signal. While such experiments can be performed in a controlled laboratory environment, limited usability exists for practical applications.

1.2.1.2 Single-Ended Algorithms

As opposed to double-ended measurement, single-ended measurement is a more recent research field. The first signal-based approach proposed in the literature dates back to 1994 [45]. In this study, comparisons between features of the received speech signal and vector quantizer codebook representations of the features of clean speech were used to estimate speech quality. In [46, 47], VQ codebooks were replaced by Gaussian mixture probability models to improve quality measurement performance. Other proposed schemes have made use of vocal tract models [48] and spectro-temporal representations of normative speech behavior [49] for single-ended quality measurement. ITU-T Recommendation P.563 represents the current state-of-the-art single-ended algorithm [50, 51]. In this thesis, the performance of the proposed algorithms are compared to the performance attained with the P.563 algorithm, hence a more detailed description of the algorithm is provided below.

The P.563 algorithm combines three principles, as depicted in Fig. 1.2 [51]. First, vocal tract and linear prediction (LP) analysis is performed to detect unnaturalness in the speech signal. The vocal tract is modeled as a series of tubes of different lengths and time-varying cross-sectional areas. From the speech signal, cross-sectional areas are evaluated for unnatural behavior. Similarly, higher-order statistics (skewness and kurtosis), computed for LP coefficients and cepstral coefficients, are investigated to

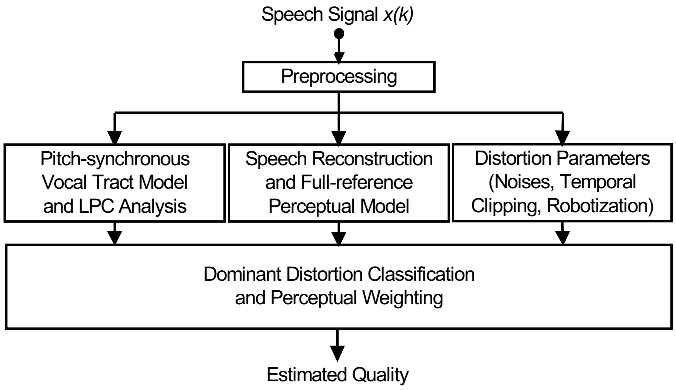


Figure 1.2: Schematic representation of ITU-T P.563.[a]

[a]Figure taken from [51, 52], with permission from the authors.

see if they lie within the restricted range expected for natural speech. Second, a pseudo-reference signal is reconstructed by modifying the computed LP coefficients to fit the vocal tract model of a typical human speaker. The pseudo-reference signal serves as input, along with the degraded speech signal, to a double-ended algorithm (similar to ITU-T P.862 [36]) to generate a "basic voice quality" measure. Lastly, specific distortions such as noise, temporal clippings, and robotization effects (voice with metallic sounds) are detected.

A total of 51 characteristic signal parameters are calculated. Based on a restricted set of eight key parameters, one of six major distortion classes is detected. The distortion classes are, in decreasing order of "annoyance": high level of background noise, signal interruptions, signal-correlated noise, speech robotization, and unnatural male and female speech [51]. For each distortion class, a subset of the extracted parameters is used to compute an intermediate quality rating. Once a major distortion class is detected, the intermediate score is linearly combined with eleven other parameters to derive a final quality estimate. While the algorithm is shown to be reliable for many telecommunications scenarios, recent research has suggested that P.563 performance is compromised for VoIP applications [38, 53, 54], noise suppressed speech [55, 56], reverberant speech [57], and synthesized speech [52], thus signaling the need for more accurate single-ended quality meters.

1.2.2 Parameter-Based Measures

For IP networks, parameter-based measures are often termed "link parametric" and make use of network parameters to estimate listening and/or conversational subjective quality, as depicted in Fig. 1.1 (c). Commonly used network parameters include codec and packet loss concealment type, packet loss pattern (random or bursty), packet loss rate, jitter, and delay. Such parameters are commonly obtained from the real-time transport protocol (RTP) header [58], real-time transport control protocol (RTCP) [59], and RTCP extended reports (RTCP-XR) [60].

Link parametric measurement was first proposed in the early 1990's by the European Telecommunications Standards Institute (ETSI). The ETSI computation model (so-called E-model) was developed as a network planning tool [61]. In the late 1990's, the E-model was standardized by the ITU-T as Recommendation G.107 [62]. Several enhanced versions of Recommendation G.107 were proposed between 2000-2005 in order to incorporate more modern transmission scenarios. Today, the E-model is a widely used transmission planning tool that describes several parametric models of specific network impairments and their interaction with subjective quality [62]. The basic assumption is that transmission impairments can be transformed into psychological impairment factors, which in turn, are additive in the psychoacoustic domain.

A transmission rating factor R is obtained from the impairment factors by

$$R = R_0 - I_s - I_d - I_{e-eff} + A, \tag{1.1}$$

where I_s, I_d, and I_{e-eff} represent speech transmission impairment factors (e.g., impairments due to quantization distortion), delay impairment factors (e.g., impairments due to echoes), and effective equipment impairment factors (e.g., impairments due to packet loss for different codec types), respectively. R_0 describes a base factor representative of the signal-to-noise ratio and A an advantage (or expectation) factor. The advantage factor serves as an offset that accounts for user expectations of the quality of service. As example, for wireline communications, $A = 0$ is used. In turn, for satellite communications in remote locations where a minimum of two satellite hops are warranted, an advantage factor of $A = 20$ is recommended [62].

The R rating ranges from 0 (bad) to 100 (excellent) and can be mapped to MOS-CQE (if the delay impairment factor I_d is considered) or MOS-LQE using equations described in ITU-T Recommendation G.107 Annex B [62]. Over the years, an extensive list of equipment impairment factors has been derived [62–64]. In addition, ITU-T Recommendations P.833 [65] and P.834 [66] have been proposed to describe methodologies used to obtain equipment impairment factor values from subjective tests and instrumental models such as PESQ, respectively.

Recently, methodologies have been proposed to compute equipment impairment factors for wideband speech codecs [67]. Moreover, as mentioned previously, the E-model is a transmission planning tool and is not recommended for online quality measurement. Hence, several extensions have been proposed to improve performance for online monitoring. It is known, for example, that the simplifying assumption that impairments are additive in the perceptual domain does not hold true for high levels of “orthogonal” (unrelated) impairments. Proprietary algorithms, such as Telchemy's VQmon, use nonlinear impairment combination models that are shown to be more accurate when high levels of dissimilar impairments are present [68].

For concatenative TTS systems, in turn, parameter-based measures are often termed “system parametric” and make use of TTS system parameters to estimate quality. A representative measure is described in [69] where an average concatenative cost function is used to assess the naturalness of the synthesized speech signal. The measure is derived from the input text and speech corpus and is inversely proportional to overall quality – the higher the number of concatenations, the lower is the quality.

Parameter-based methods have gained wide popularity due to their reduced computational complexity. As an example, studies have suggested that the E-model (link parametric) can be up to 1000 times less computationally complex, in terms of millions of instructions per second, than P.563 (signal-based) [70]. As will be shown in Chapter 4, the performance of parameter-based methods can be compromised if signal-based distortions (e.g., temporal clippings) are present since such distortions are not captured by connection and/or system parameters. These shortcomings motivate the need for a hybrid signal-and-parameter-based methodology. Previously proposed hybrid architectures are described next.

1.2.3 Hybrid Measurement - Previous Investigations

Hybrid signal-and-link-parametric measurement methods use link parameters in addition to the voice payload to estimate subjective quality, as illustrated with Fig. 1.1 (d). A few hybrid approaches have been proposed previously. In [71, 72], PESQ is used to estimate the quality of the received speech signal and the estimated MOS-LQO is converted into an equipment impairment factor which, along with transmission delay estimates, is input to the E-model. While such approaches are useful to quickly obtain non-tabulated equipment impairment factors, the high computational complexity, the need for a clean reference signal, and the sensitivity of PESQ to connection parameters make them impractical for online quality of service (QoS) control. Moreover, the use of PESQ for systems equipped with noise suppression algorithms is not recommended [40], thus limiting its usability for emerging wireless-VoIP communications.

More recently, the work described in [73] proposes a hybrid methodology where temporal clippings and signal-to-noise ratio (SNR) are estimated from the degraded speech signal in a single-ended manner using advanced signal processing techniques. Real-time transport protocol (RTP) and real-time transport control protocol (RTCP) analysis is used to obtain the packet loss rate. Different impairment models are computed and combined with the E-model for a final quality rating. The algorithm is shown to correlate well with PESQ quality scores; however, due to the aforementioned PESQ limitations with VoIP speech data, it is not obvious if the method accurately predicts subjective quality. Moreover, the performance and complexity of the hybrid scheme is not compared to benchmark algorithms such as the E-model and P.563, thus its improvement over existing algorithms is not clear.

1.3 Motivation and Objectives

In this thesis, we investigate the development of innovative *single-ended* quality measures for emerging speech communication applications. Focus is placed on low-complexity methods that allow for measurement of multiple quality dimensions. Motivations and objectives for subsequent chapters are detailed below.

1.3.1 General-Purpose Quality Measurement

General-purpose objective quality measures constitute the perfect candidate for online quality monitoring and control. Single-ended quality “probes” can be distributed at different points throughout a network to pinpoint locations where different quality degradations occur, thus allowing for specific corrective measures to be taken. Moreover, today, speech signals are transmitted over multi-stage hybrid networks and are exposed to a plethora of different sources of distortion. Such heterogeneous processing motivates degradation-type identification which enables the deployment of appropriate corrective measures to assure that QoS remains at acceptable levels.

The objective is to develop techniques to detect and measure multiple distortion sources such as multiplicative noise and temporal discontinuities. Furthermore, current state-of-the-art quality measurement algorithms use complex signal processing techniques, during *online* operation, to estimate quality. In order to reduce algorithm computational complexity, we seek an alternate avenue where the majority of the processing is performed *offline*. Our objective is to use statistical models of normative speech behaviour, obtained during offline training, to detect distortions and to quantify their effects on perceived speech quality. In Chapter 2, Gaussian mixture statistical models are investigated and key algorithmic operational modules, many of which are adapted and applied in subsequent chapters, are introduced.

1.3.2 Quality Measurement for Noise-Suppressed Speech

With advances in speech communication technologies, noise suppression has become essential for applications such as hearing aids, mobile phones, and voice-controlled systems. Noise suppression, however, can introduce unwanted perceptual artifacts such as "musical noise" [20]. Current objective quality measurement algorithms are shown to perform poorly for noisy speech processed by noise suppression algorithms [74, 75]. Hence, to date, a generally accepted evaluation metric for noise suppressed speech is *not* available. This limitation has motivated the search for quality measurement algorithms that are tuned to noise suppressed speech.

The objective is to develop an algorithm to measure the three quality dimensions described in [19], namely, signal distortion, background intrusiveness, and overall quality. Moreover, as emphasized in [76], unwanted noise can be suppressed at three different stages in the speech transmission chain: prior to speech coding, in the network, or at the decoder. Given the rise in emerging heterogeneous networks and transcoding scenarios, we seek to develop a *distributed* quality diagnosis tool that allows for reliable detection of noise suppression processing and for accurate prediction of noise-suppressed speech quality. Chapter 3 focuses on the development of quality diagnosis tools for noise suppressed speech signals.

1.3.3 Quality Measurement for Wireless-VoIP Communications

VoIP has increased in popularity over the past few years, mainly due to its low cost and capability of integrating data and real-time voice traffic on existing network infrastructures. As mentioned previously, with VoIP communications, objective quality measurement can be performed either on a signal basis or a link parametric basis. While signal-based algorithms perform well for traditional telephony applications, algorithm performance has been shown to decrease when applied to VoIP communications [37, 38, 53, 54]. Link parametric approaches, in turn, can be severely affected by distortions that are not captured by connection parameters. Such sensitivity poses a serious threat to emerging wireless-VoIP communications, which are expected to become ubiquitous in the near future [4, 5]. With wireless-VoIP, speech signals can be corrupted by varying levels and types of background noise prior to packetization. Moreover, in more advanced wireless communications algorithms, noise suppression artifacts can also be introduced.

The objective here is to develop an algorithm that overcomes the limitations of both pure signal-based and pure link parametric quality measurement. As such, a hybrid signal-and-link-parametric approach to single-ended quality measurement of packet speech is proposed in Chapter 4. The method makes use of IP connection parameters to determine a base quality representative of the packet transmission network. Signal-based distortions, resulting from the signal processing in the wireless communications chain, are then detected and quantified from the speech signal and used to adjust the base quality accordingly.

1.3.4 Quality Measurement for Hands-Free Communications

With the advances in far-field hands-free communication technologies, signal processing algorithms have been developed to combat unwanted reverberation effects. With reverberant speech, objective measures computed from the *measured* room impulse response (IR), such as reverberation time (T_{60}) and direct-to-reverberation energy ratio (DRR), are often used to characterize signal quality. Offline measurement of room impulse responses, however, is a laborious task. In addition, the impulse response varies with acoustic source positioning, room temperature, as well as placement of room furnishings. As a consequence, room acoustical parameters obtained from room

IR measurements are not feasible for real-time signal processing applications. Moreover, with dereverberated speech, room impulse responses need to be *estimated* (e.g., via blind deconvolution) and often result in poor quality characterization. Dereverberation algorithms are also known to introduce audible artifacts to the speech signal and such artifacts are not captured by the estimated room IR. These shortcomings motivate the need for a *signal-based* quality measure.

The objective here is two-fold: (1) to develop signal-based measures to "blindly" characterize room acoustics (e.g., estimate T_{60} and DRR) and (2) to develop a single-ended speech quality measure for reverberant and dereverberated speech. With environment-sensitive systems such as automatic speech/speaker recognition, blind source separation, and pitch tracking algorithms, signal-based measures can be used for systematic parameter adaptation to best match room acoustical properties. In Chapter 5, speech temporal dynamics information is used to characterize room acoustical parameters. Objective quality measures are also developed for three quality dimensions, namely, colouration, reverberation tail effects, and overall quality.

1.3.5 Quality Measurement for Text-to-Speech Systems

Applications involving text-to-speech (TTS) systems are emerging continuously. In the past, applications served mostly as an aid to the visually impaired. Today, TTS systems are also being applied in email and short message service readers, automated directory assistance, foreign language education, and assistive and augmentative communications, to name a few applications. Evaluation of synthesized speech is not an easy task as various quality dimensions need to be assessed (e.g., naturalness, intelligibility). Moreover, with synthesized speech, clean reference signals are often not available, thus limiting the use of double-ended objective quality measures. To date, a blind perceptual quality estimator for synthesized speech does *not* exist and the work described in Chapter 6 attempts to bridge this gap. The goal is to use hidden Markov models (HMM), trained on naturally produced speech, as artificial reference models with which synthesized speech signals are assessed. The temporal information captured by the HMM allows for accurate estimation of several quality dimensions including *overall impression*, *naturalness* and *continuity/fluency*.

1.4 Thesis Contributions

The aim of this dissertation is to develop objective quality measurement algorithms for emerging speech communications and applications. The key contributions are:

1. The development of a general-purpose speech quality measurement algorithm based on statistical models of normative speech behaviour. Innovative methods to detect and quantify distortions caused by multiplicative noise and temporal discontinuities are also developed. A slightly modified version of the proposed temporal discontinuity detection algorithm has since been incorporated into the ANIQUE+ algorithm developed by Alcatel-Lucent [77, 78]. Moreover, the multiplicative noise detection/quantification algorithm has been incorporated into the Deutsche Telekom's proprietary spoken dialogue evaluation system described in [79]. Publications that have resulted from this contribution include [46, 47, 74, 80–82].

2. The proposal of two objective quality measures for noise-suppressed speech. The first is a *network-distributed* measurement algorithm which subsumes current

single- and double-ended architectures [74]. The approach allows for double-ended measurement without the need of a clean reference signal. With the proposed architecture, it is possible to analyze the quality of the system under test and both quality degradations and quality enhancements can be detected and handled; such functionality is not available with existing methods. The second is a single-ended measure and makes use of statistical reference models of clean, noisy, and noise-suppressed speech [55]. Kullback-Leibler distances, computed between online trained models and offline obtained reference models, are proposed as indicators of speech quality. Models are developed for active and inactive speech segments, thus allowing for measurement of three quality dimensions, namely, signal distortion, background intrusiveness, and overall quality. Publications that have resulted from this contribution include [55, 74].

3. Analysis of variance tests are conducted to investigate the performance of current state-of-the-art signal-based and link parametric objective measures for burgeoning wireless-VoIP communications [83]. Signal-based schemes are shown to be sensitive to VoIP network-related transmission parameters. Link parametric schemes, in turn, are severely affected by distortions that are not captured by the connection parameters; representative degradations may include acoustic background noise, temporal clippings, and noise suppression artifacts. To overcome this limitation, a hybrid signal-and-link-parametric quality measurement algorithm is proposed. A codec-integrated methodology is further proposed to allow for "feature sharing" between the speech codec and the quality measurement algorithm. Under such an integrated configuration, the proposed scheme has processing time that is approximately 90% lower than that attained with

the current state-of-the-art single-ended algorithm ITU-T P.563. Publications that have resulted from this contribution include [53, 84].

4. The proposal of a reverberation-to-speech modulation energy ratio measure for blind characterization of room acoustics and for single-ended quality measurement of reverberant and dereverberated speech. The proposed measure is computed from a spectro-temporal signal representation where speech and reverberation tail components are shown to be separable [85]. Signal-based estimators of the room reverberation time and direct-to-reverberation energy ratio parameters are devised, the latter being the first of its kind. An adaptive measure is also introduced and shown to be useful for objective measurement of reverberant and dereverberated speech. The proposed measure allows for estimation of multiple quality dimensions, namely, colouration and reverberation tail effects, and overall quality. Publications that have resulted from this contribution include [57, 85, 86].

5. The first steps towards the development of a reference-free objective quality measure for synthesized speech are taken. Hidden Markov models, trained on naturally produced speech, serve as artificial text- and speaker-independent reference models with which synthesized speech signals are assessed. A normalized log-likelihood measure, computed between perceptual features extracted from synthesized speech and a gender-dependent reference model, is proposed and shown to be a reliable measure for multidimensional TTS quality diagnosis. The proposed measure allows for accurate estimation of quality dimensions labeled *overall impression, listening effort, naturalness, continuity/fluency*, and *acceptance*. This contribution has resulted in the publication of [87].

1.5 Thesis Organization

This thesis is based on a collection of eight manuscripts ([46, 55, 74, 83–87]). Some minor modifications to the papers are made, mostly to provide common notation and to remove repetitive introductory material. It is emphasized, however, that there is still some overlap in content, mostly in the description of the features, the statistical models, and the performance figures used; this overlap, however, should assist the reader to follow the development of the thesis. In Chapter 2, the general-purpose objective quality measurement algorithm is introduced and described. Many of the components and signal processing modules described in this chapter are modified and used in subsequent chapters. Chapter 3 focuses on quality measurement of noise suppressed speech, while hybrid signal-and-link-parametric measurement is addressed in Chapter 4. Quality measurement for reverberant and dereverberated speech signals is described in Chapter 5 and for synthesized speech signals in Chapter 6. Lastly, Chapters 7 and 8 provide a general discussion and the conclusions, respectively.

Chapter 2

General-Purpose Objective Speech Quality Measurement

2.1 Preamble

This chapter is compiled from material extracted from manuscripts published in the IEEE Transactions of Audio, Speech, and Language Processing [74] and IEEE Signal Processing Letters [46]. Earlier versions appeared in the Proceedings of the 2005 and 2006 International Conference on Acoustics, Speech, and Signal Processing [47, 80].

2.2 Introduction

Despite all of the advances in modern telecommunication networks, subjective speech quality measurement has remained costly and labor intensive. For the purpose of real-time speech quality measurement on a network-wide scale, low complexity general-purpose objective speech quality estimation is needed. In this chapter, one such quality meter is presented. The proposed algorithm is constructed from models of speech signals, including clean and degraded speech, and speech corrupted by multiplicative noise and temporal discontinuities. Machine learning methods are used to design the models, including Gaussian mixture models, support vector machines, and random forest classifiers. Estimates of the subjective mean opinion score (MOS-LQS) generated by the models are combined using hard or soft decisions generated by a classifier which has learned to match the input signal with the models.

The remainder of this chapter is organized as follows. In Section 2.3, a detailed description of the single-ended algorithm is given. Algorithm design considerations are covered in Section 2.4 and algorithm performance is evaluated in Section 2.5. Conclusions are reported in Section 2.6.

2.3 Overview of the Proposed Algorithm

In the proposed method, single-ended measurement algorithms are designed based on the architecture depicted in Fig. 2.1. Perceptual features are first extracted from the test speech signal every 10 milliseconds. The time segmentation module labels the feature vector of each frame as belonging to one of three possible classes: active-voiced, active-unvoiced, or inactive (background noise). Signals are then processed by a multiplicative noise detector. During design, the detector is optimized in conjunction with the "noise estimation and MOS mapping" and the "consistency calculation and MOS mapping" modules. A preliminary quality score, namely $MOS_{tmp,1}$, is computed from the estimated amount of multiplicative noise present in the signal. A second preliminary score, $MOS_{tmp,2}$, is computed from six consistency measures, which in turn, are calculated relative to reference models of speech behaviour.

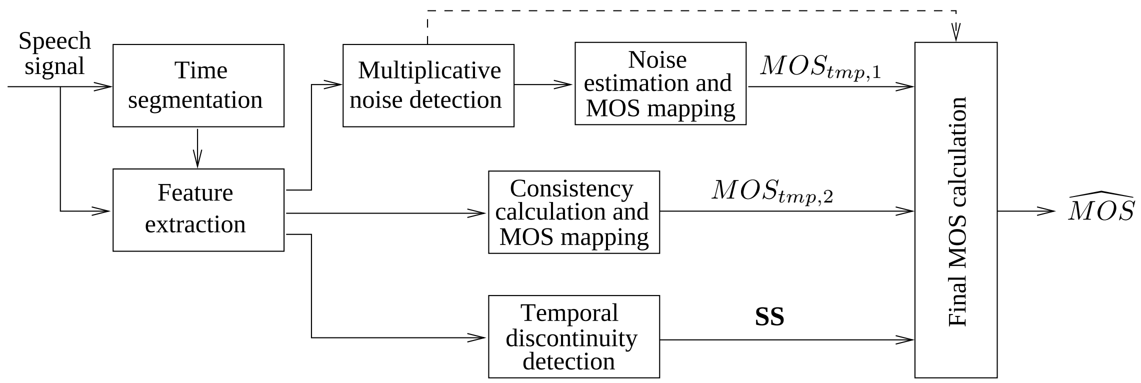


Figure 2.1: Architecture of the proposed general-purpose single-ended measurement algorithm.

It is noted that $MOS_{tmp,1}$ provides more accurate speech quality estimates, relative to $MOS_{tmp,2}$, for certain degradation conditions. The objective of the multiplicative noise detector is, thus, to distinguish which conditions can be better represented by $MOS_{tmp,1}$. Lastly, temporal discontinuities (**SS**) are detected and a final quality rating ($\widehat{MOS}$) is computed. The final rating is a linear combination of the preliminary scores adjusted by the negative effects that temporal discontinuities have on perceived quality. A detailed description of each block is provided in the remainder of this section. Experimental optimization of algorithm parameters is presented in Section 2.4.2.

2.3.1 Time Segmentation and Feature Extraction

Time segmentation is employed to separate the speech frames into different classes. It has been shown that each class exerts different influence on the overall speech quality [46]. Time segmentation is performed using a voice activity detector (VAD) and a voicing detector. The VAD identifies each 10-millisecond speech frame as being active or inactive (background noise). The voicing detector further labels active frames as voiced or unvoiced. The VAD from the adaptive multi-rate (AMR) speech codec [88] (VAD option 1) and the voicing determination algorithm described in [89] are used.

Perceptual linear prediction (PLP) cepstral coefficients [90] serve as primary features and are extracted from the speech signal every 10 milliseconds. The coefficients are obtained from an "auditory spectrum," constructed to exploit three essential psychoacoustic precepts. First, the spectrum of the original signal is warped into the Bark frequency scale and a critical band masking curve is convolved with the signal. The signal is then pre-emphasized by a simulated equal-loudness curve to match the frequency magnitude response of the ear. Lastly, the amplitude is compressed by the cubic-root to match the nonlinear relation between intensity of sound and perceived loudness. The auditory spectrum is then approximated by an all-pole autoregressive model, whose coefficients are transformed to P^{th} order PLP cepstral coefficients $\mathbf{x} = \{x_i\}_{i=0}^{P}$; the zeroth cepstral coefficient x_0 is employed as an energy measure [91]. When describing the PLP vector for a given frame m, the notation $\mathbf{x}_m = \{x_{i,m}\}_{i=0}^{P}$ is used. Moreover, the PLP vector averaged over N_f frames ($\bar{\mathbf{x}}$) is given by

$$\bar{\mathbf{x}} = \frac{1}{N_f} \sum_{m=1}^{N_f} \mathbf{x}_m. \tag{2.1}$$

The order of the autoregressive model determines the amount of detail in the auditory spectrum preserved by the model. Higher order models tend to preserve more speaker-dependent information and are more complex to calculate. We experiment with 5^{th} and 10^{th} order PLP coefficients. On our databases both models incur similar quality estimation performance; thus, for the benefit of lower computational complexity, 5^{th} order PLP coefficients are chosen. Fifth order models have been successfully used in [45] and are shown in [90] to serve well as speaker-independent speech spectral parameters. Moreover, dynamic features in the form of delta and double-delta coefficients [91] have been shown to indicate the rate of change (speed) and the acceleration of the spectral components, respectively [92]. As will be shown in Section 2.3.4, the delta information for the zeroth PLP cepstral coefficient can be used to detect temporal discontinuities.

Lastly, the mean cepstral deviation ($\bar{\sigma}$) of a test signal is computed. In Section 2.3.2, it will be shown that $\bar{\sigma}$ can be used to detect and estimate the amount of multiplicative noise. The mean cepstral deviation is the average of all "per-frame" deviations (σ_m) of the PLP cepstral coefficients (excluding the zeroth coefficient). The per-frame deviation is defined as

$$\sigma_m = \sqrt{\frac{1}{P-1}\sum_{i=1}^{P}\left(x_{i,m} - \left(\frac{1}{P}\sum_{j=1}^{P} x_{j,m}\right)\right)^2} \tag{2.2}$$

and $P = 5$.

2.3.2 Detecting and Estimating Multiplicative Noise

It is known that multiplicative noise (also known as speech-correlated noise) can be introduced by logarithmically companded PCM (e.g., G.711) or ADPCM (e.g., G.726) systems as well as by other waveform speech coders [93]. In fact, the modulated noise reference unit (MNRU) [94] was originally devised to reproduce the perceptual distortion of log-PCM waveform coding techniques. MNRU systems produce speech that is corrupted by controlled speech-amplitude-correlated noise. The speech plus multiplicative noise output, $y_{MNRU}(n)$, of an MNRU system is given by

$$y_{MNRU}(n) = v(n) + v(n)10^{-Q/20}N(n), \tag{2.3}$$

where $v(n)$ is the clean speech signal and $N(n)$ is white Gaussian noise (unit variance). The amount of multiplicative noise, $v(n)10^{-Q/20}N(n)$, is controlled by the parameter

Q, which represents the ratio of input speech power to multiplicative noise power, and is expressed in decibels (dB). This parameter is often termed the "Q value".

Measuring multiplicative noise of the form (2.3), when *both* the clean signal and the degraded speech signals are available, is fairly straightforward. The task becomes more challenging when the original clean signal is unavailable. In such instances, Q must be estimated. To the best of our knowledge, the scheme presented in [50] is the only published method of estimating multiplicative noise using only the degraded speech signal. The process entails an evaluation of the spectral statistics of the signal during active speech periods.

Today, MNRU degradations and reference waveform codecs such as G.711 and G.726 are used extensively as "anchor" conditions in testing and standardization of emerging codec technologies and in network planning. Current speech quality measurement algorithms should handle such degradation conditions efficiently. In previous work [80], estimating multiplicative noise is shown to be beneficial for GMM-based speech quality measurement. A multiplicative noise estimator, similar to the one described in [50], was deployed and performance improvement was reported for MNRU degradations. This improvement in performance substantiates the need for an efficient method of estimating multiplicative noise. Here, an innovative and simple technique is employed.

The technique is based on PLP coefficients and their mean cepstral deviations. As discussed in [95], the multiplicative noise term in (2.3) introduces a fairly flat noise floor in regions of the spectrum of $y_{MNRU}(n)$ where the power of $v(n)$ is small. On the other hand, in regions where the power of the input signal is sufficiently large, the spectrum of $v(n)$ is almost perfectly preserved (see examples in [95]). The

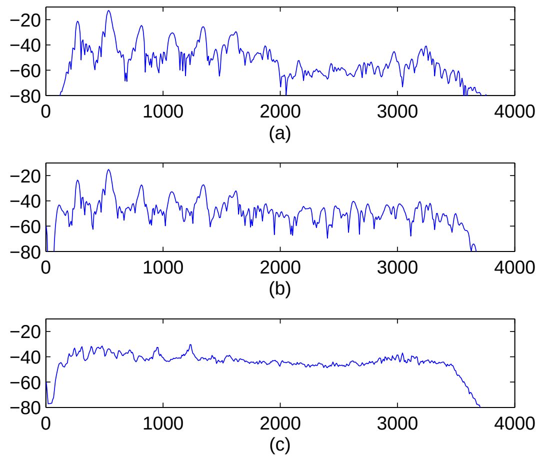


Figure 2.2: Spectrum of a speech frame (a) before processing, and after (b) 25 dB MNRU and (c) 5 dB MNRU processing. The x-axis represents frequencies in Hz and the y-axis amplitudes in dB.

amount of multiplicative noise is controlled by the parameter Q. As a result, as Q approaches 0 dB (i.e., power of multiplicative noise equals the power of input speech), the flat spectral characteristic of the multiplicative noise starts to dominate the spectrum of $y_{MNRU}(n)$. In such instances, information about the spectral envelope of the signal is lost, deteriorating the quality and intelligibility of the signal. To illustrate this behavior, Fig. 2.2 (a)-(c) shows the spectrum of a speech frame prior to processing and after MNRU degradation with $Q = 25$dB, and $Q = 5$dB, respectively. As can be clearly seen, the spectrum of $y_{MNRU}(n)$ becomes flatter as the amount of multiplicative noise increases (i.e., as Q decreases).

The use of mean cepstral deviation as a measure of the amount of multiplicative noise present in a signal is inspired by the definition of cepstrum – the inverse Fourier transform of the log-spectrum of a signal [91]. Tests on our databases show that the cepstral deviation for MNRU speech correlates well with the flatness of the log-spectrum, i.e., with the amount of multiplicative noise. As an example, a correlation of -0.93 is attained between the mean cepstral deviation of active speech frames ($\bar{\sigma}_{active}$) and Q values for MNRU-degraded speech files on our speech databases. Negative correlation is expected since lower Q values result in flatter spectra. In turn, spectrum and cepstrum are related via a Fourier transformation, thus a flat spectrum translates into a non-flat cepstrum, i.e., a high $\bar{\sigma}_{active}$. Once Q is estimated, $MOS_{tmp,1}$ can be computed via simple regression. In fact, a polynomial mapping can be employed directly between $\bar{\sigma}_{active}$ and $MOS_{tmp,1}$. As will be shown in Section 2.4.2, $MOS_{tmp,1}$ provides accurate estimates of perceived subjective quality for various different degradation conditions, in addition to corruption by MNRU multiplicative noise.

In this chapter, the detection of the presence of high levels of multiplicative noise is treated as a supervised classification problem. In fact, the detector is trained to detect not only multiplicative noise, but also all other degradation conditions where $MOS_{tmp,1}$ is better than $MOS_{tmp,2}$ as an estimator of MOS-LQS (some example conditions are given in Section 2.4.2.3). Detection is performed on a "per-signal" basis and depends on a 14-dimensional input consisting of the PLP vector averaged over active frames ($\bar{\mathbf{x}}_{active}$) and over inactive frames ($\bar{\mathbf{x}}_{inactive}$), and the mean cepstral deviation for active frames ($\bar{\sigma}_{active}$) and for inactive frames ($\bar{\sigma}_{inactive}$). Inactive frames are used as they provide cues for discriminating additive background noise from speech-correlated noise. Experiments are carried out with support vector classifiers (SVC) [96], classification and regression trees (CART) [97], and random forest classifiers (RFC) [98] as candidate detectors. Training of the detectors will be described in more detail in Section 2.4.2.3.

2.3.3 GMMs, Consistency Calculation and MOS Mapping

Gaussian mixture models (GMMs) are used to model the PLP cepstral coefficients of each of the three classes of speech frames – voiced, unvoiced, and inactive. A Gaussian mixture density is a weighted sum of M component densities

$$p(\mathbf{u}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{u}), \tag{2.4}$$

where $\alpha_i \geq 0, i = 1, ..., M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{u})$ are $\mathcal{K}$-variate Gaussian densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The parameter list, $\boldsymbol{\lambda}$=$\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$, defines a particular Gaussian mixture density, where $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i\}$. The well-known expectation-maximization (EM) algorithm [99] is used to iteratively estimate $\boldsymbol{\lambda}$ from training data.

In pilot experiments it has been found that accuracy can be enhanced if the algorithm is also equipped with information regarding the behavior of speech degraded by different transmission and coding schemes [80]. To this end, clean speech signals are used to train three different Gaussian mixture densities, $p_{clean,class}(\mathbf{u}|\boldsymbol{\lambda})$. The subscript "class" represents either voiced, unvoiced, or inactive frames. For the degradation model, $p_{degraded,class}(\mathbf{u}|\boldsymbol{\lambda})$ are trained.

For the benefit of low computational complexity, we make a simplifying assumption that vectors between frames are independent. This assumption has been shown in [46] to provide accurate speech quality estimates. Nonetheless, improved performance is expected from more sophisticated models, such as hidden Markov models, where statistical dependencies between frames can be considered. This investigation, however, is left for future study. Thus, for a given speech signal, the consistency between the observation and the models is defined as the normalized (log-)likelihood

$$c_{model,class}(\mathbf{X}_{class}) = \frac{1}{N_{class}} \sum_{j=1}^{N_{class}} \log(p_{model,class}(\mathbf{x}_{class,j}|\boldsymbol{\lambda})) \tag{2.5}$$

where $\mathbf{X}_{class} = \{\mathbf{x}_{class,i}\}_{i=1}^{N_{class}}$ denotes the set of all N_{class} PLP vectors that have been classified as belonging to a given speech *class*. The subscript "model" represents either the clean or the degradation reference model. Normalization is required as N_{class} varies for different test signals.

In total, six consistency measures are calculated per test signal. For each class, the product of the consistency measure (2.5) and the fraction of frames of that class in the speech signal is computed; this product is referred to as a "feature." In the rare case when the fraction of frames of a specific class is zero (e.g., only voiced speech is detected), a constant $c_{model,class} = c = -15$ is used as the feature. Lastly, the six features are mapped to $MOS_{tmp,2}$. We experiment with multivariate polynomial regression and multivariate adaptive regression spline (MARS) [100] as candidate mapping functions. With MARS, the mapping is constructed as a weighted sum of truncated linear functions (see [33] for more detail). On our databases, MARS is shown to provide superior performance. MARS models are designed based on the MOS-LQS of degraded speech. Simulation results show that a simple MARS function composed of a linear combination of 18 truncated linear functions provides accurate quality estimation performance. The experimental results presented in Section 2.5 make use of a MARS model to map the 6-dimensional consistency feature vector, calculated on a per-signal basis, into $MOS_{tmp,2}$.

2.3.4 Temporal Discontinuity Detection

Motivated by the results reported in [101] and by first and second order methods used for edge detection in images (e.g., [102, 103]), we employ delta and double-delta coefficients for temporal discontinuity detection. Delta coefficients represent the local time derivatives (slope) of the cepstral sequence and are computed according to

$$\Delta \mathbf{x}_m = \sum_{l=-L}^{L} l \ \mathbf{x}_{m+l}, \tag{2.6}$$

where the normalization factor $\sum_{l=-L}^{L} l^2$ is omitted as it does not affect the simulation results. Delta coefficients indicate the rate of change (speed) of spectral components; in our simulations $L = 5$ is used. Double-delta coefficients are the second-order local time derivatives of the cepstral sequence and are computed according to

$$\Delta^2 \mathbf{x}_m = \sum_{n=-N}^{N} n \Delta \mathbf{x}_{m+n}. \tag{2.7}$$

Double-delta coefficients indicate the acceleration of the spectral components; in our simulations $N = 3$ is used.

As mentioned in Section 2.3.1, the zeroth PLP cepstral coefficient is used as an energy term. The delta and double-delta features, calculated from x_0, provide insight into the dynamics of the signal energy. The main assumption used here is that for natural speech, abrupt changes in signal energy do not occur. The two main temporal impairments that should be detected are abrupt starts and abrupt stops [49]. In abrupt starts, the signal energy, its rate of change, and acceleration increase abruptly. The opposite occurs with abrupt stops. This behavior is illustrated with Figs. 2.3 and 2.4. In Fig. 2.3 (a)-(d), the waveform of a speech signal, the energy, and energy rate of change (termed Δ_0 for simplicity) and acceleration (termed Δ_0^2) are

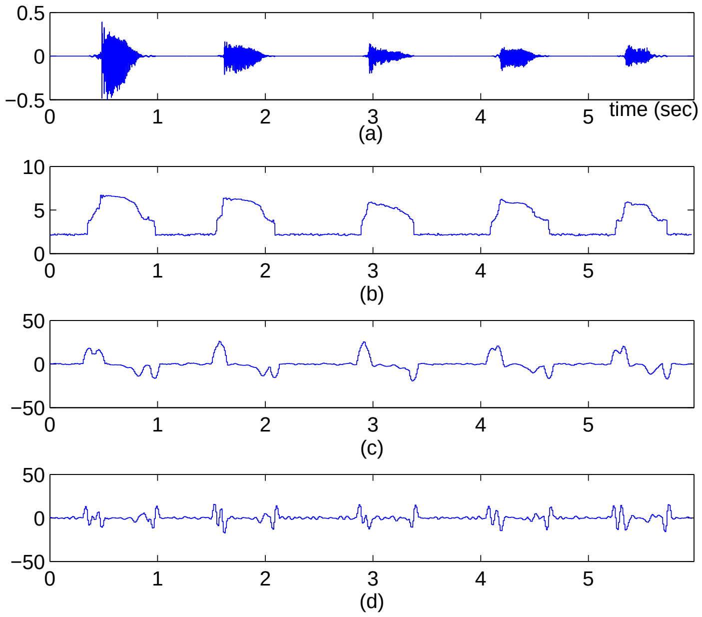


Figure 2.3: Analysis of a signal's (a) waveform, (b) x_0, (c) Δ_0, and (d) Δ_0^2. Signal consists of five vowels uttered in a noisy office environment.

depicted, respectively. The signal consists of five vowels uttered by a male speaker in a noisy office environment. Vowels are chosen as their extremities are often erroneously detected as abrupt starts or stops. Notice the subtle spikes in Δ_0 and Δ_0^2 at each vowel extremity. In Fig. 2.4, temporal discontinuities, or "clippings," have been introduced at the beginning or at the end of each vowel. The abrupt starts and stops are indicated with arrows. Notice that the unnatural changes cause abnormal spikes in Δ_0 and Δ_0^2.

To detect abrupt starts or stops, two steps are required. First, the energy of frame at time t_c is compared to the energy of frame $t_c+\tau$. If the energy increase (or decrease) surpasses a certain threshold T, then a candidate abrupt start (or stop) is detected. The parameters T and τ are optimized on our training database, as described in Section 2.4.2.4. Once a candidate discontinuity is detected, a support vector classifier is used to decide whether in fact a temporal discontinuity has occurred. The SVC

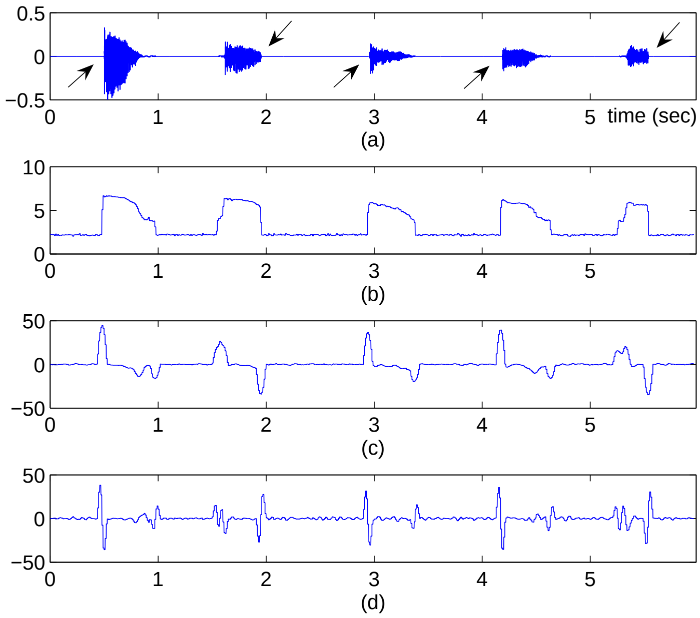


Figure 2.4: Analysis of a "clipped" signal's (a) waveform, (b) x_0, (c) Δ_0, and (d) Δ_0^2. Abrupt starts and stops are indicated with arrows.

is only invoked at candidate discontinuities in order to reduce the computational complexity of the algorithm. Here, two SVCs are used: one tests for abrupt starts (given a sudden increase in x_0) and the other for abrupt stops (given a sudden decrease in x_0). Input features to the SVC are Δ_0 and Δ_0^2 for the $\tau + F$ frames preceding t_c and the $\tau + F$ frames succeeding t_c. The parameter F is empirically set to 2, resulting in a 10-dimensional input feature vector. The output of each classifier is one of two possible classes, namely, "discontinuity" or "non-discontinuity." We experiment with linear, polynomial and radial basis function (RBF) support vector classifiers; on our databases, an RBF SVC attained superior performance.

The output **SS** of the temporal discontinuity detection block, as depicted in Fig. 2.1, is a $(n_b + n_s + 2)$-dimensional vector comprised of the number of detected abrupt starts (n_b) and abrupt stops (n_s) and the approximate time at which each discontinuity occurs. As an example, suppose for a given speech file three abrupt starts are detected at times $\mathbf{t}_b = \{t_{b1}, t_{b2}, t_{b3}\}$ and two abrupt stops at times $\mathbf{t}_s = \{t_{s1}, t_{s2}\}$. The resulting parameter $\mathbf{SS}$ is represented by $\mathbf{SS} = \{n_b, \mathbf{t}_b, n_s, \mathbf{t}_s\}$.

2.3.5 Final MOS Calculation

The final MOS-LQO calculation is based on a linear combination of the intermediate MOSs, adjusted by the negative effects temporal discontinuities have on perceived quality, i.e.,

$$\widehat{MOS} = p_m MOS_{tmp,1} + (1 - p_m) MOS_{tmp,2} - C(\mathbf{SS}). \tag{2.8}$$

Here, p_m is the probability that $MOS_{tmp,1}$ is better than $MOS_{tmp,2}$ as an estimator of MOS-LQS. This statistic is calculated by the detector on a "per-signal" basis. More detail regarding the computation of p_m is given in Section 2.4.2.5.

The term $C(\mathbf{SS})$ resembles the effects temporal discontinuities have on perceived quality. Experiments, such as [104], suggest that humans can perform continuous assessment of time-varying speech quality. It is also noted that the location of a discontinuity within a signal can affect the listener's perception of quality; this short-term memory effect is termed "the recency effect." Impairments detected at the end of the signal have more negative effect on the perceived quality than impairments detected at the beginning. In [49], a decay model is used to emulate the recency effect. More recently, however, experiments carried out in [105] suggest that the recency effect is harder to observe in speech signals of short time duration. Instead, a "subconscious integration" is performed where unconsciously, multiple degradations are combined and reported as a single level of speech quality.

Since the files in our databases are of short time durations (on average 8 seconds) we do not consider the recency effect and model $C(\mathbf{SS})$ as

$$C(\mathbf{SS}) = C(n_b, n_s) = n_b K_b + n_s K_s \tag{2.9}$$

where K_b and K_s are penalty terms for the detected abrupt starts and stops, respectively. These constants are optimized on the training databases, as will be discussed in Section 2.4.2.5. In this chapter, since the recency effect is not considered, $\mathbf{t}_s$ and $\mathbf{t}_b$ are not computed. Nevertheless, for longer speech files, (2.9) can be modified to incorporate such temporal information; in particular, a decay model can be employed.

2.4 Algorithm Design Considerations

In this section, algorithm design considerations are described in detail.

2.4.1 Database Description

In total, 20 MOS-LQS labeled databases are used in our experiments. The speech databases are described in Table 2.1. We separate fourteen databases for training (databases 1-14) and the remaining six are used for testing (databases 15-20). Additionally, during training several algorithm parameters need to be optimized. To this end, 20% of the training set is randomly chosen to be used for parameter validation; henceforth, this subset will be referred to as the "validation set." Parameter calibration is discussed in further detail in Section 2.4.2. The content of each database is described next.

Table 2.1: Properties of speech databases used in our experiments.

Database	Language	No. of Files	No. of Conditions	Training	Testing
1	French	176	44	✓	
2	Japanese	176	44	✓	
3	English	176	44	✓	
4	French	200	50	✓	
5	Italian	200	50	✓	
6	Japanese	200	50	✓	
7	English	200	50	✓	
8	English	96	24	✓	
9	English	96	24	✓	
10	English	240	60	✓	
11	Italian	2440	20	✓	
12	Japanese	2440	20	✓	
13	English	2440	20	✓	
14	English	2088	46	✓	
15	English	3072	48		✓
16	English	3072	48		✓
17	English	3072	48		✓
18	English	3328	52		✓
19	English	96	24		✓
20	English	448	28		✓

Databases 1-7 are the ITU-T P-series Supplement 23 (Experiments 1 and 3) multilingual databases [106]. The three databases in Experiment 1 have speech processed by various codecs (G.726, G.728, G.729, GSM-FR, IS-54 and JDC-HR), singly or in different cross-tandem configurations (e.g., G.729–G.728–GSM-FR). The four databases in Experiment 3 contain single- and multiple-encoded G.729 speech under various channel error conditions (0-10% bit error rate and 0-5% random and burst frame erasure rate) and input noise conditions (clean, vehicle, street, and hoth noises).

Databases 8 and 9 are two wireless databases with speech processed, respectively, by the IS-96A and IS-127 EVRC (Enhanced Variable Rate Codec) codecs under various channel error conditions (forward and reverse 3% frame erasure rate) with or without the G.728 codec in tandem. Database 10 is a mixed wireless-wireline database with speech under a wide range of degradation conditions – tandemings, channel errors, temporal clippings, and amplitude variations. A more detailed description of the conditions in database 10 can be found in [107]. Databases 11-13 comprise speech coded using the G.711, G.726 and the G.728 speech coders, alone and in various different tandem configurations. Database 14 has speech from standard speech coders (G.711, G.726, G.728, G.729, and G.723.1), under various channel degradation conditions (clean, 0.01% bit error rate, and 1-3% frame erasure rate).

Databases 15-17 comprise speech coded with the 3GPP2 Selectable Mode Vocoder (SMV) under different tandeming, channel impairments, and environment noise degradation conditions. Database 18 has speech from standard speech coders (G.711, G.726, G.728, G.729E, and GSM-EFR) and speech processed by a cable VoIP speech coder, under various channel degradation conditions. Lastly, databases 19 and 20 have speech recorded from an actual telephone connection in the San Francisco area and live network speech samples collected from AMPS, TDMA, CDMA, and IS-136 forward and reverse links. In all databases described above, speech degraded by different levels of MNRU is also included.

Speech files from databases 15-20 are used solely for testing and are unseen to the algorithm. Databases 15-18 are kept for testing as they provide speech files coded using newer codecs than the codecs represented in the training datasets. Evaluation using these databases demonstrates the applicability of the proposed algorithm to emerging codec technologies. Database 19 has speech files that are composed of two spoken utterances, one by a male speaker and the other by a female speaker, thus are regarded as being composite male-female signals. Although this is not common in listening tests, we are interested in seeing how robust the proposed algorithm is to speaker and gender changes. Furthermore, database 20 is composed of speech files that have been processed by older wireless codecs. Many of the files in this database are of poor speech quality ($MOS < 2$) and comprise degradation conditions not represented in the training datasets.

2.4.2 Algorithm Parameter Calibration

In order to optimize algorithm parameters, preliminary "calibration" experiments are carried out. In the sequel, we describe the steps taken to calibrate each of the processing blocks depicted in Fig. 2.1.

2.4.2.1 Multiplicative Noise Estimation and MOS Mapping

For optimization of the multiplicative noise estimator, MNRU degraded training files are used. Experiments are carried out with 2^{nd} and 3^{rd} order polynomial mappings between $\bar{\sigma}_{active}$ and the Q value. On the validation set, the latter presented better performance. The estimated amount of multiplicative noise achieved a 0.92 correlation with the true Q value. For comparison, the multiplicative noise estimator described in [50] resulted in a correlation of 0.66. For the noise-to-MOS mapping, it is found that a simple linear regression between the estimated amount of multiplicative noise and $MOS_{tmp,1}$ suffices. The two mappings are replaced by one single 3^{rd} order polynomial mapping between $\bar{\sigma}_{active}$ and $MOS_{tmp,1}$. A 0.95 correlation between $MOS_{tmp,1}$ and the true MOS-LQS is attained for MNRU validation files.

2.4.2.2 Consistency Calculation

To calibrate the consistency calculation block, an effective combination of GMM configuration parameters (M and covariance matrix type) needs to be found. For voiced and unvoiced frames we experiment with diagonal matrices and M=8, 16, or 32, and M=2, 3, or 5 for full covariance matrices. For inactive frames, we only experiment with diagonal matrices and M=2, 3, or 6. The calibration experiment suggests the use of 3 full GMM components for voiced frames and 32 diagonal components for unvoiced frames, for both the clean and the degradation model. For inactive frames, 6 diagonal components are needed for the degradation model and 3 for the clean model. This is consistent with the fact that for clean speech, inactive frames have virtually no signal energy and fewer Gaussian components are required.

The consistency-to-MOS mapping is designed using a MARS regression function with parameters optimized using degraded MOS-LQS labeled training files. The function maps the six consistency measures into $MOS_{tmp,2}$. As mentioned previously, the designed MARS regression function is composed of a simple weighted sum of 18 truncated linear functions. The mapping is performed once per speech signal and incurs negligible computational complexity (approximately 18 scalar multiplications and 54 scalar additions). For files in the validation set, a 0.82 correlation is attained between $MOS_{tmp,2}$ and the actual MOS-LQS; if MNRU degraded files are removed, the correlation increases to 0.86. This result suggests that a combination of $MOS_{tmp,1}$ and $MOS_{tmp,2}$ may lead to better performance when compared to using $MOS_{tmp,2}$ alone.

2.4.2.3 Multiplicative Noise Detection

The multiplicative noise detector is optimized to select the best preliminary quality score, $MOS_{tmp,1}$ or $MOS_{tmp,2}$, for a given test signal. To gain a sense of which conditions are best represented by each preliminary score, tests are performed on the training set where the true MOS-LQS is known. As expected, of 288 files processed by the G.711 and G.726 codecs, 252 are better represented by $MOS_{tmp,1}$. Similarly, of 252 MNRU-degraded files with 0 dB$< Q <$ 35 dB, 209 are better represented by $MOS_{tmp,1}$. If only files with $Q <$ 20 dB are considered, 103 (out of 108) are better estimated by $MOS_{tmp,1}$. The primary objective of the detector, thus, is to detect signals corrupted by high levels of multiplicative noise.

Nonetheless, for some degradation conditions other than multiplicative noise conditions, $MOS_{tmp,1}$ is also shown to be a better estimator of MOS-LQS than $MOS_{tmp,2}$. Some examples include speech signals processed by low bitrate vocoders (e.g., G.723.1 at 5.3 kbit/s), where the quality of five (out of 32) of the signals is better represented by $MOS_{tmp,1}$. Moreover, of 112 samples processed by medium bitrate codecs (e.g., G.729E at 11.8 kbit/s), the quality of 22 signals is better estimated by $MOS_{tmp,1}$

than by $MOS_{tmp,2}$. As a consequence, in instances where high levels of multiplicative noise are not detected, the classifier learns which temporary score results in the best estimation performance.

To calibrate the detector, first, all training samples are processed by the top *and* middle branches depicted in the block diagram in Fig. 2.1. The estimated preliminary MOSs are compared to the true MOS-LQS and all samples in which the top branch achieved the smallest estimation error receive a label "TOP"; otherwise a label "MID" is assigned. This new labeled training set determines which preliminary score best estimates the true MOS-LQS for a given speech signal and is used to train the detector. The detector can be designed to operate in two different modes: hard-decision or soft-decision. In hard-decision mode, the detector selects the single best preliminary quality score and $p_m \in \{0, 1\}$ is used in (2.8). With this mode only one preliminary score needs to be computed. On the contrary, soft-decision detection requires that both preliminary scores be estimated, and a "weight" is assigned to each score. The weight ($0 \leq p_m \leq 1$) is computed by the detector on a "per-signal" basis and reflects the probability of $MOS_{tmp,1}$ more accurately predicting MOS-LQS than $MOS_{tmp,2}$. The term p_m resembles the likelihood of the presence of high levels of multiplicative noise in the signal. After detector optimization, signals in the validation set with high levels of multiplicative noise have p_m that approach unity.

We experiment with three different candidate classifiers: CART, SVC and RFC. The classifiers are trained using the aforementioned labeled training set. An RFC is an ensemble of unpruned decision trees induced from bootstrap samples of the training data. The final class decision is based on a majority vote from all individual trees (see [98] for more details regarding random forest classifiers). On our validation set, an RFC with 500 trees achieved the best classification performance; all files with high levels of multiplicative noise (e.g., MNRU with $Q < 12$ dB) were correctly detected.

2.4.2.4 Temporal Discontinuity Detection

Calibrating the temporal discontinuity detector encompasses the determination of parameters T and τ, and training of the support vector classifiers. On our data it was found that if the values of x_0 doubled (or halved) within 20-50 milliseconds a candidate discontinuity could be detected. With these possible values of τ, the SVCs correctly identified all abrupt stops and starts on the validation dataset. In an attempt to reduce the number of times the SVCs are executed, a more stringent threshold, $\tau = 2$ (equivalent to 20 milliseconds), is used.

2.4.2.5 Final MOS Calculation

Lastly, the parameters in (2.8) are optimized. Initially, $C(\mathbf{SS}) = 0$ is assumed and we experiment with hard-decision detection and soft-decision detection. On the validation set, soft-decision detection resulted in superior performance. With soft-decision detection, p_m is computed by the RFC and represents the fraction of the 500 individual decision trees that have selected $MOS_{tmp,1}$ as the best estimator of subjective quality. Once the soft-decision mode is set, the parameters K_b and K_s in (2.9) are estimated by minimizing the squared error between (2.8) and the true MOS-LQS for "clipped" training signals. On our data, $K_b = 0.09$ and $K_s = 0.13$ were found. These parameters are consistent with [49], where it is argued that the abrupt stops have, intuitively, a more significant impact on perceived speech quality relative to abrupt starts.

2.5 Test Results

In this section we compare the proposed algorithm to P.563 using the test databases described in Section 2.4.1. The performance of the algorithms is assessed by the Pearson correlation (R) between the N MOS-LQS (w_i) and MOS-LQO (y_i) samples,

$$R = \frac{\sum_{i=1}^{N}(w_i - \bar{w})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(w_i - \bar{w})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{2.10}$$

where $\bar{w}$ is the average of w_i, and $\bar{y}$ is the average of y_i. MOS measurement accuracy is assessed using the root-mean-square error ($RMSE$),

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(w_i - y_i)^2}{N}}. \tag{2.11}$$

Table 2.2 presents "per-condition" R and $RMSE$ between condition-averaged MOS-LQS and condition-averaged MOS-LQO, for each of the test datasets. The results are obtained after an individual 3^{rd} order monotonic polynomial regression for each dataset, as recommended in [50]. The column labeled "$\%R \uparrow$" lists the percentage "R-improvement" obtained by using the proposed GMM-based method over P.563. The R-improvement is given by

$$\%R \uparrow = \frac{R_{GMM} - R_{P.563}}{1 - R_{P.563}} \times 100\% \tag{2.12}$$

and indicates percentage reduction of P.563's performance gap to perfect correlation. The column labeled "$\%RMSE$" lists percentage reduction in $RMSE$, relative to P.563, by using the proposed scheme. As can be seen, the proposed algorithm outperforms P.563 on all test databases. An average R-improvement of 44% and an average reduction in $RMSE$ of 17% is attained.

An interesting result is obtained with database 19. Recall that this database had MOS-LQS labeled speech signals composed of two utterances, one spoken by a male

Table 2.2: Performance comparison on unseen test datasets. Results are per-condition after 3^{rd} order polynomial regression.

Test Database	P.563		Proposed			
	R	$RMSE$	R	$\%R\uparrow$	$RMSE$	$\%RMSE$
15	0.863	0.253	0.908	32.8	0.206	18.6
16	0.835	0.274	0.864	17.6	0.249	9.1
17	0.748	0.273	0.868	47.6	0.212	22.3
18	0.916	0.218	0.939	27.4	0.187	14.2
19	0.421	0.456	0.868	77.2	0.455	0.2
20	0.758	0.569	0.909	62.4	0.362	36.4
Average	–	–	–	**44.2**	–	**16.8**

speaker and the other by a female speaker. On this database, P.563 achieves a poor correlation of 0.421. In fact, before applying the 3^{rd} order monotonic polynomial mapping, P.563 achieves a very poor $R = 0.121$. This may be due to the fact that P.563 depends on vocal tract analysis to test for unnaturalness of speech. By rating the unnaturalness of speech separately for male and female voices, P.563 is compromised for composite male-female signals. As a sanity check, we test the performance of PESQ (with the mapping described in [108]) and an $R = 0.974$ and $RMSE = 0.422$ is attained.

The plots in Fig. 2.5 show MOS-LQO versus MOS-LQS for the proposed algorithm and for P.563. Each data point represents one of the 332 different degradation conditions available in the test databases. In these plots and in the performance figures described below the composite male-female quality estimates are left out. Plots (a) and (b) illustrate the relationship between GMM MOS-LQO and MOS-LQS, before and after 3^{rd} order monotonic polynomial regression (optimized on each test dataset), respectively. Prior to polynomial mapping, an overall $R = 0.874$ and $RMSE = 0.321$

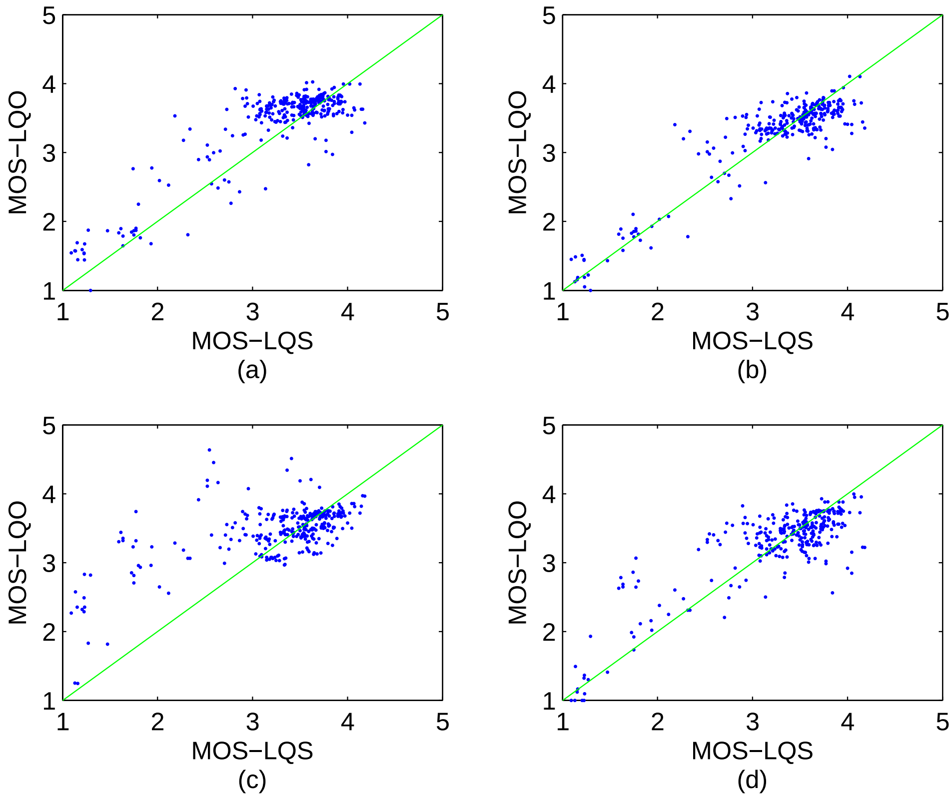


Figure 2.5: Per-condition MOS-LQO versus MOS-LQS for (a) proposed algorithm prior to and (b) after 3^{rd} order monotonic polynomial mapping, and for (c) P.563 before and (d) after the polynomial mapping.

is attained; after the mapping, $R = 0.918$ and $RMSE = 0.221$. Similarly, plots (c) and (d) illustrate the relationship between P.563 MOS-LQO and MOS-LQS, before and after the monotonic mapping. An overall $R = 0.7281$ and $RMSE = 0.391$ is attained prior to regression; after regression $R = 0.853$ and $RMSE = 0.292$.

The 3^{rd} order monotonic polynomial regression is suggested in [50] in order to map the objective score onto the subjective scale. This mapping is used to compensate for variations of the MOS-LQS scale across different subjective tests, variations due to different voter groups, languages, contexts, amongst other factors. Monotonic mappings perform scale adjustments but do not alter the ranking of the objective scores. Ultimately, the goal in objective quality estimation is to design algorithms whose quality scores rank similarly to subjective quality scores. This is due to the fact that objective scores that offer good ranking performance produce accurate MOS-LQS estimates, given a suitable monotonic mapping is used for scale adjustment. To this end, we use rank-order correlations as an additional figure of merit of algorithm performance. Rank-order correlations are calculated using (2.10), with the original data values replaced by the *ranks* of the data values; this measure is often termed Spearman's rank correlation coefficient (R_S). For P.563, a "per-condition" $R_S = 0.705$ is attained on the test data. The proposed algorithm achieves $R_S = 0.793$, a 30% R-improvement. The results presented above, for all three performance measures, suggest that the proposed algorithm provides more accurate estimates of subjective quality relative to the current state-of-the-art P.563 algorithm.

2.6 Conclusions

This chapter has described the development of a *general-purpose* single-ended speech quality estimation algorithm. The algorithm employs speech signal models designed using machine learning methods. Innovative methods to detect and quantify multiplicative noise and temporal distortions are described. Comparisons with the current state-of-the-art P.563 algorithm demonstrate the efficacy of the algorithm and its potential for providing more accurate quality measurements. The GMM-based quality measurement paradigm and the algorithmic signal processing modules described in this chapter serve as the foundation for many of the algorithms proposed in subsequent chapters.

Chapter 3

Quality Measurement for Noise-Suppressed Speech

3.1 Preamble

This chapter is compiled from material extracted from manuscripts published in the IEEE Transactions of Audio, Speech, and Language Processing [74] and the Journal of Multimedia [55].

3.2 Introduction

With the advances in speech communication technologies, noise suppression has become essential for applications such as hearing aids, mobile phones, and voice-controlled systems. In the past, various double-ended measures were proposed to characterize the performance of noise suppression algorithms (e.g., see [109]). Such measures, however, did not take into account human perceptual characteristics, thus, did not correlate well with subjective quality. More recently, widely used double-ended objective measures were tested as quality estimators of noise suppressed speech [75, 110, 111]; low correlations with subjective quality were reported for most measures. Included in the measures was the current state-of-the-art double-ended ITU-T PESQ algorithm. In fact, to date the only blind estimator suitable for noise suppressed speech is the ITU-T standard algorithm P.563 [50]. Experiments described herein, however, suggest that low correlations are attained with subjective quality [74]; results reported in [56] corroborate such findings. In this chapter, two methods are proposed for reliable objective quality measurement of noise suppressed speech (Sections 3.3-3.4); conclusions are presented in Section 3.5.

The first method subsumes both single-ended and double-ended quality measurement paradigms. Single-ended quality probes are distributed along the transmission chain and used to detect *where* in the network noise suppression is performed. Side information, sent from probes situated in the chain prior to and post noise suppression, is used by a quality diagnosis module to characterize the performance of the noise suppression algorithm. Moreover, once noise suppression is detected, multiple quality dimensions – signal distortion, background intrusiveness, and overall quality – are blindly estimated. The proposed architecture allows for both quality degradations and quality enhancements to be detected and handled, a functionality that is not available with existing double- or single-ended algorithms. The second method allows for low complexity blind estimation of perceptual quality and is based on statistical reference models of clean, noisy, and noise-suppressed speech. Kullback-Leibler distances, computed between online trained models and offline obtained reference models, are used as indicators of multiple speech quality dimensions.

3.3 Distributed Quality Measurement

Existing single- and double-ended algorithms are only capable of estimating the quality of the received signal per se. In order to analyze the quality of a transmission system, assumptions on the input signal are needed. As mentioned in Section 1.2.1.1, double-ended algorithms presuppose that the input is undistorted. Moreover, it is assumed that the output is of quality no better than the input. Current double-ended algorithms would fail if any of these assumptions were to fail.

A scenario where *both* assumptions are not met can be seen in Fig. 3.1 where a clean signal x_{clean} suffers impairments that degrade speech quality. Common impairments may include interference on an analog access network, environment noise, noise introduced by equipment within the network, and lost packets in a VoIP network. The noisy signal x_{noisy} is then input to a speech enhancement system and the enhanced output $x_{enhance}$ is of quality better than x_{noisy}. Such a system configuration commonly occurs when using a noise reduction algorithm to enhance speech. As will be shown in Section 3.3.2.2, the performance of current double-ended schemes may be compromised when only x_{noisy} and $x_{enhance}$ are made available to the algorithm.

3.3.1 Measurement Configuration

The objective is to devise a measurement scheme that subsumes current single- and double-ended measurement architectures. The approach allows for double-ended measurement *without* the underlying assumptions that the input signal needs to be clean and that the output needs to be of quality better than the input. With the proposed architecture, it is possible to analyze the quality of the system under test and both quality degradations and quality enhancements can be detected and handled. This

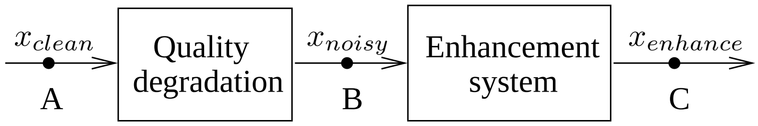


Figure 3.1: Block diagram of a speech enhancement system.

section will give emphasis to *quality enhancements*, in particular to noise suppression.

The proposed architecture is depicted in Fig. 3.2. The conventional double-ended algorithm is replaced by two single-ended schemes, one at the input and another at the output of the system being tested, and a system diagnosis tool. This configuration requires access to the output signal and to information extracted from the input signal, thus can be viewed as a "reduced-reference" measurement paradigm. Since the amount of information extracted from the input signal is negligible (in terms of bitrate), the proposed scheme is much more economical than existing double-ended architectures which require access to the actual input signal.

In analogy to Fig. 3.1, if the input single-ended algorithm is placed at the point labeled "A" and the output single-ended algorithm is placed at the point labeled "B" then quality degradations are handled. On the other hand, if the input single-ended algorithm is placed at the point labeled "B" and the output single-ended algorithm is placed at the point labeled "C" then quality enhancements are handled. Here, focus will be placed on the latter scenario as it represents the case where the input signal is *not* clean and the output *is* of quality better than the input. As mentioned previously, the performance of current double-ended schemes may be compromised in this scenario.

To allow for accurate speech quality measurement of noise suppressed signals, the proposed GMM-based algorithm described in Section 2.3 is updated to incorporate

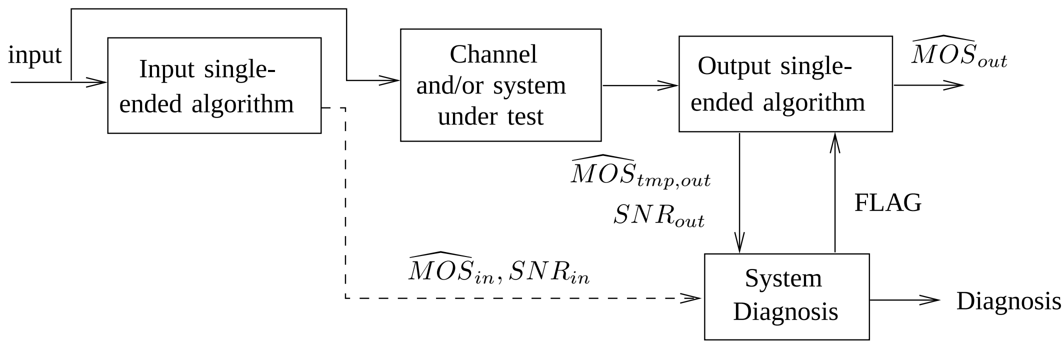


Figure 3.2: Architecture of the proposed distributed quality measurement paradigm.

reference models of noise suppressed speech signals. Similar to the clean and degradation models, the "noise-suppressed" model is designed for voiced, unvoiced, and inactive frames. If noise suppression is detected, consistency measures relative to all three reference models (clean, degraded and noise suppressed) are computed; otherwise, consistency measures are computed only for the clean and degradation models. In the latter case, the MARS mapping described in Section 2.4.2.2 is used; in the former, a separate MARS mapping is trained on a subjectively scored noise-suppressed speech database. Details regarding the database will be given in Section 3.3.2. Noise-suppressed reference model and MARS mapping design considerations will be given in Section 3.3.2.1.

With the proposed architecture, the transmission of input measurements (this is illustrated with the dashed arrow in Fig. 3.2) and a system diagnosis module are necessary in order to detect if noise suppression has occurred. We have investigated the effectiveness of transmitting the input SNR (SNR_{in}), computed by the VAD algorithm, and the input MOS-LQO ($\widehat{MOS}_{in}$), estimated based on the consistency measures calculated relative to the clean and the degradation reference models. As mentioned previously, the amount of side information is negligible.

At the output end, SNR_{out} is computed and a preliminary MOS-LQO ($\widehat{MOS}_{tmp,out}$) is estimated based on the clean and the degradation model-based consistency measures. These two measures are sent to the system diagnosis module. The diagnosis module, in turn, sends a flag back to the output single-ended algorithm indicating whether noise suppression has been detected. Detection occurs if $SNR_{in} < SNR_{out}$ and $\widehat{MOS}_{in} + \gamma < \widehat{MOS}_{tmp,out}$, where γ is the standard deviation of the estimated input MOS-LQO ($\gamma = 0.29$ on the noise suppressed database). With this detection rule all of the noise suppressed speech files were correctly detected.

If noise suppression is detected, the output single-ended algorithm calculates a final MOS-LQO ($\widehat{MOS}_{out}$) based on consistency measures calculated relative to the three reference models, otherwise $\widehat{MOS}_{out} = \widehat{MOS}_{tmp,out}$. Other diagnostic tests, such as measuring (in terms of MOS) the amount of quality degradation (or enhancement) imparted by the transmission system, or measuring SNR improvement, are also possible. Further characterization of the noise suppression algorithm may be aided with the transmission of other input measurements (e.g., see measures in [112]).

3.3.2 Experimental Results

The proposed architecture is tested using the subjectively scored NOIZEUS database [113]. The database comprises speech corrupted by four types of noise (babble, car, street, and train) at two SNR levels (5 and 10dB) and processed by 13 different noise suppression algorithms; a total of 1792 speech files are available. The noise suppression algorithms fall under four different classes: spectral subtractive, subspace, statistical-model based, and Wiener algorithms. A complete description of the algorithms can be found in [21, 113].

The subjective evaluation of the NOIZEUS database was performed according to ITU-T Recommendation P.835 [19]. As mentioned in Section 1.1.2, with the P.835 methodology, listeners are instructed to successively attend to and rate three different signal components of the noise suppressed speech signal: (1) the speech signal alone and (2) the background noise alone using the scales described in Table 1.4, and (3) the overall speech-plus-noise content using the ACR scale shown in Table 1.1. The average scores over all listeners are termed SIG-LQS, BCK-LQS, and OVRL-LQS, respectively; OVRL is equivalent to the MOS-LQS described in [14].

3.3.2.1 Design Considerations

In order to train reference models of noise suppressed speech signals and to design the updated MARS mapping function, the NOIZEUS database has to be separated into a training and a test set. We perform this separation in three different ways to test the robustness of the proposed architecture to different unseen test conditions. First, speech files are separated according to noise levels; files with SNR = 10dB are used for training and files with SNR = 5dB are left for testing. Second, speech signals are separated according to noise sources. Signals corrupted by street and train noise are used for training and signals corrupted by babble and car noise are left for testing. Lastly, speech files are separated according to noise suppression algorithms. For training, noisy signals processed by spectral subtractive and subspace algorithms are used; noisy signals processed by statistical-model based and Wiener algorithms are left for testing. The number of conditions for each of the three test sets described above are 52, 52, and 64, respectively, out of a total of 104 degradation conditions for the entire database.

To design the reference models for noise suppressed speech signals we experiment with different combinations of GMM parameters. It is observed that for all three tests 32 diagonal components for voiced and unvoiced frames and 6 diagonal components for inactive frames strike a balance between accuracy and complexity. Moreover, a separate MARS mapping function is designed for each of the three tests. Each MARS function maps a 9-dimensional feature vector into $\widehat{MOS}_{out}$.

3.3.2.2 Test Results

In this section, we compare the performance of the proposed architecture to that of PESQ. Two different PESQ configurations are tested: (1) a hypothetical configuration where the original clean signal is available, and (2) a more realistic scenario where only the noisy and the noise suppressed signals are available. Configuration 1 makes use of the clean signal as reference input and although evaluation of noise reduction systems is not recommended, as described in [40], the results to follow suggest accurate estimation performance. On the other hand, configuration 2 exemplifies the case where the reference input signal is not clean, and the quality of the output is better than that of the input. As will be shown in the sequel, this configuration compromises PESQ performance. Moreover, swapping the input signals (i.e., noise suppressed signal to reference input and noisy signal to degraded input) brought no improvement.

Table 3.1 presents "per-condition" R and $RMSE$ between condition-averaged OVRL-LQS and condition-averaged OVRL-LQO, for the three test sets. Results are reported after 3^{rd} order monotonic polynomial regression (for PESQ the mapping proposed in [108] is not used as it degrades performance substantially). As can be seen, when the original clean speech signal is available, PESQ achieves accurate

Table 3.1: Performance comparison with PESQ and P.563 on the three test sets. Configuration 1 makes use of the original clean signal and the noise-suppressed signal, configuration 2 of the noisy signal and the noise-suppressed signal.

Test No.	PESQ - Config. 1		PESQ - Config. 2		Proposed		P.563	
	R	$RMSE$	R	$RMSE$	R	$RMSE$	R	$RMSE$
1	0.886	0.178	0.610	0.305	0.861	0.196	0.587	0.311
2	0.864	0.233	0.581	0.377	0.827	0.261	0.563	0.384
3	0.922	0.181	0.731	0.321	0.817	0.270	0.637	0.361

estimation performance. However, when only the noisy signal is available as reference, substantial improvement is attained with the proposed architecture. For comparison, Table 3.1 also shows the performance of P.563 on the three tests.

3.3.2.3 Component Quality Estimation

Existing objective measurement algorithms can only attempt to estimate OVRL-LQS. However, it is unknown how humans integrate the individual contributions of speech and noise distortions when judging the overall quality of a noise suppressed signal. To this end, devising an algorithm capable of also estimating SIG-LQS and BCK-LQS would be invaluable. The estimates can be used to test newer generations of noise reduction algorithms and to assess the algorithms' capability of maintaining speech signal naturalness whilst reducing background noise to nonintrusive levels. In [110], the NOIZEUS database is used to evaluate six double-ended objective estimates of SIG-LQS and BCK-LQS. The study makes use of the original clean signal as a reference and low correlations with subjective quality were reported ($R < 0.65$).

Due to the modular architecture of the proposed GMM-based algorithm, a simple extension can be implemented to allow for single-ended SIG-LQS and BCK-LQS

Table 3.2: Performance of SIG-LQO and BCK-LQO estimated by the proposed algorithm.

Test No.	SIG-LQO		BCK-LQO	
	R	$RMSE$	R	$RMSE$
1	0.813	0.295	0.717	0.235
2	0.804	0.355	0.728	0.289
3	0.807	0.331	0.707	0.305

estimation. In particular, two new MARS mapping functions are optimized on the training datasets. To estimate SIG-LQS, a 6-dimensional MARS function is devised to map consistency measures of voiced and unvoiced frames (for all three reference models - clean, degraded, and noise suppressed) into SIG-LQO. To estimate BCK-LQS, a simple 4-dimensional MARS function is designed to map consistency measures of inactive frames (for all three models) and the estimated SNR into BCK-LQO. Table 3.2 presents "per-condition" R and $RMSE$ between condition-averaged SIG-LQS (BCK-LQS) and condition-averaged SIG-LQO (BCK-LQO), for the three aforementioned test sets. Results are reported after 3^{rd} order monotonic polynomial regression optimized on each test set. The results are encouraging given that the original clean signal is *not* available as a reference. Next, blind estimation of noise suppressed speech quality is addressed.

3.4 Single-Ended Quality Measurement

In this section, the behavior of the PLP cepstrum is investigated for speech corrupted by additive background noise as well as noisy speech processed by a noise suppression algorithm. The obtained insights are used to develop an algorithm for blind estimation

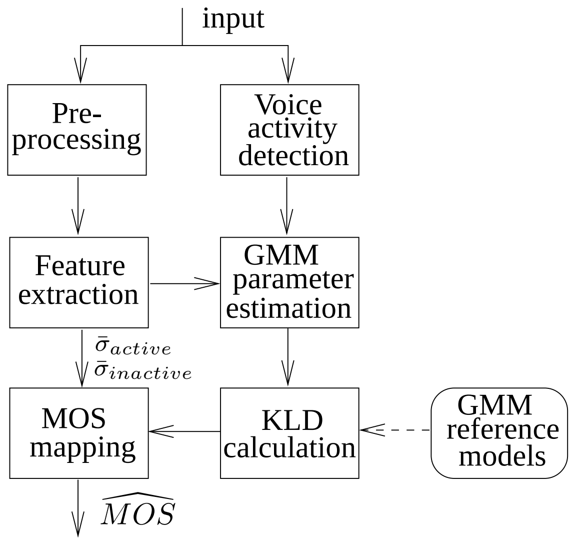


Figure 3.3: Architecture of the proposed single-ended algorithm for noise suppressed speech.

of noise suppressed speech quality, as described next.

3.4.1 Architecture of Proposed Algorithm

The overall architecture of the proposed algorithm is depicted in Fig. 3.3. First, the level of the speech signal is normalized and the signal is filtered to simulate the handsets used in listening tests. Perceptual features are then extracted from the test speech signal every 10 milliseconds. The voice activity detector (VAD) labels the feature vector of each frame as either active or inactive (background noise). Offline, three reference models are created. High-quality undistorted speech signals, signals corrupted by additive noise at low signal-to-noise ratios (SNR), and noise suppressed speech signals are used to produce reference models of the behavior of clean, noisy, and noise suppressed speech features, respectively.

In all cases, the probability distribution of the features is modeled with a Gaussian mixture model; separate models are trained for active and for inactive frames. Online, the expectation-maximization algorithm is used to estimate a GMM for the features extracted from the test signal. To achieve low-complexity processing, an approximation of the Kullback-Leibler distance (KLD) is used. KLDs are computed between the online estimated models and the three reference models. The calculated distances, together with a spectral flatness measure, serve as speech quality indicators and are mapped to an estimated mean opinion score, $\widehat{MOS}$ [19]. A detailed description of each block is provided in the remainder of this section.

3.4.1.1 Pre-processing and VAD

The pre-processing module performs level normalization and intermediate reference system (IRS) filtering. The level of the speech signal is normalized to -26 dBov using the P.56 speech voltmeter [114] and the modified IRS filter is applied to emulate the characteristic of the handset used in the listening tests (see description in [36]). Voice activity detection is employed to label speech frames as active or inactive. Recall that in Chapter 2 and in Section 3.3 a voicing detector was used to further label active frames as "voiced" or "unvoiced." Here, voicing decision is not carried out as the extra processing did not garner substantial improvement in estimation performance. The VAD from the adaptive multi-rate (AMR) speech codec is used [88].

3.4.1.2 Feature Extraction

Pilot experiments are carried out to investigate the behaviour of PLP cepstra under additive background noise (a similar experiment is described in [115] for linear prediction coefficients) and different noise suppression conditions. The plots in

Fig. 3.4 (a)-(d) illustrate the behavior observed for clean, noisy, and noise suppressed speech. The coefficients depicted in Fig. 3.4 are averaged over one thousand active speech frames. It is observed that PLP cepstral coefficients lie in distinct areas of the cepstral vector space with lower quality speech (e.g., SNR=5 dB in Fig. 3.4) lying further away from the clean speech "centroid," represented by "$\times$" in the figure. As can be seen, similar trends are found for all PLP cepstral coefficients. It is also noted that different "distances" are obtained for different noise reduction algorithms (not illustrated in the figure) and different noise levels; such insights suggest that KLDs can be advantageously used for quality measurement of noise suppressed speech.

Moreover, our experiments, in addition to the results reported in [76, 116], have suggested that a spectral flatness measure can be used to assist in clean, noisy, and noise-suppressed speech discrimination. As such, the mean cepstral deviation $\bar{\sigma}$ of the test signal, shown in Section 2.3.1 to be related to spectral flatness, is computed according to (2.2). In our experiments, $\bar{\sigma}$ is calculated for active and inactive frames separately ($\bar{\sigma}_{active}$ and $\bar{\sigma}_{inactive}$, respectively).

3.4.1.3 Reference GMMs and Parameter Estimation

Gaussian mixture models are used to model the PLP cepstral coefficients of active and of inactive speech frames. Gaussian mixture densities are given by (2.4) and are described in more detail in Section 2.3.3. In subsequent sections, a Gaussian mixture density will be represented by $\boldsymbol{\lambda}$=$\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$, where $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i\}$. *Offline*, six Gaussian mixture densities, $p_{model,class}(\mathbf{x}|\boldsymbol{\lambda})$ are trained. The subscript "model" represents either clean, noisy, or noise suppressed; the subscript "class" represents either active or inactive frames. *Online*, the EM algorithm [99] is used to train a GMM

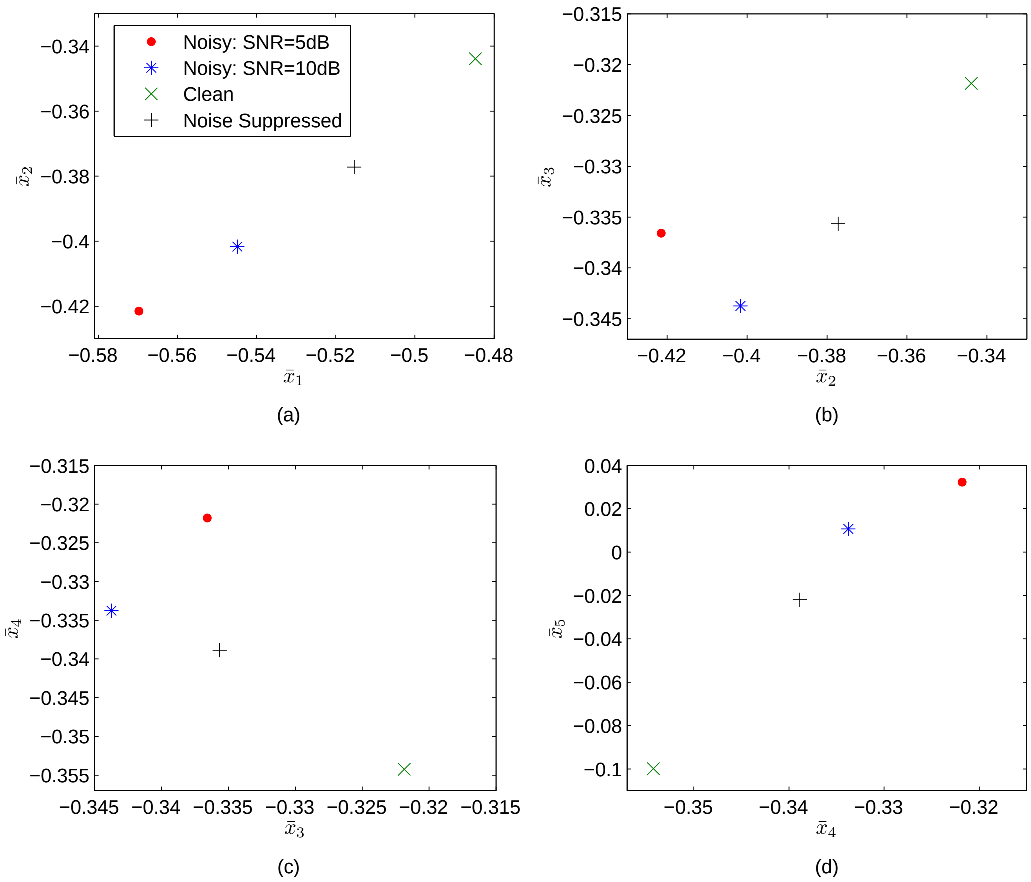


Figure 3.4: PLP cepstral behavior for clean speech ($\times$), speech corrupted by background noise with an SNR of 5 dB ($\bullet$) and 10 dB ($\star$), and noisy speech processed by a noise reduction algorithm ($+$). Cepstral coefficients are averaged over 1000 active speech frames. The plots depict (a) $\bar{x}_2$ versus $\bar{x}_1$, (b) $\bar{x}_3$ versus $\bar{x}_2$, (c) $\bar{x}_4$ versus $\bar{x}_3$, (d) $\bar{x}_5$ versus $\bar{x}_4$.

on features extracted from the test signal; a separate model is found for active and for inactive frames ($\tilde{p}_{class}(\mathbf{x}|\tilde{\boldsymbol{\lambda}})$). Pilot experiments show that if the EM algorithm is initialized using the *k-means* algorithm it converges in approximately 17 iterations for the active models and in 7 iterations for the inactive models. Alternate initialization schemes were also tested but resulted in no significant performance improvement.

3.4.1.4 KLD Calculation and MOS Mapping

The Kullback-Leibler distance measures the "distance" between two probability density functions $p_1(x)$ and $p_2(x)$ by

$$D(p_1, p_2) = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} dx. \tag{3.1}$$

$D(p_1, p_2)$ describes how well $p_2(x)$ approximates $p_1(x)$. The KLD is calculated between the online-estimated model ($\tilde{p}$ for short) and the three reference models (p), for active and inactive frames. Commonly, the Monte Carlo method is used to compute the integral in (3.1); this, however, is prohibitively expensive for online quality measurement. We experiment with two fast approximations of the KLD; one (termed $D1$) assumes equal number of Gaussian components between reference and test models $M = \tilde{M}$ [117], while the other ($D2$) allows for $M \neq \tilde{M}$ [118]. $D1$ is given by

$$D1(p, \tilde{p}) = \sum_{i=1}^{M} \alpha_i \log \frac{\alpha_i}{\tilde{\alpha}_i} + \sum_{i=1}^{M} \alpha_i \; D(b_i(\mathbf{x}), \tilde{b}_i(\mathbf{x})), \tag{3.2}$$

where

$$\begin{aligned} D(b_i(\mathbf{x}), \tilde{b}_i(\mathbf{x})) = \frac{1}{2}\Big(& \log\Big(\frac{\det \tilde{\boldsymbol{\Sigma}}_i}{\det \boldsymbol{\Sigma}_i}\Big) + \mathrm{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\Sigma}_i) \\ & + (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T \tilde{\boldsymbol{\Sigma}}_i^{-1} (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) - \mathcal{K}\Big) \end{aligned} \tag{3.3}$$

is the KLD between two $\mathcal{K}$-variate Gaussian densities.

The approximation described in [118] is posed as a linear programming problem. While many algorithms are available to solve the problem efficiently, they are often complex and time consuming. For the sake of reduced computational complexity, a simplification is performed and $D2$ is defined here as

$$D2(p, \tilde{p}) = \sum_{i=1}^{M} \sum_{j=1}^{\tilde{M}} \alpha_i \tilde{\alpha}_j D(b_i(\mathbf{x}), \tilde{b}_j(\mathbf{x})). \tag{3.4}$$

Note from (3.1)-(3.4) that $D1$ and $D2$ are asymmetric measures, i.e., $D(p, \tilde{p}) \neq D(\tilde{p}, p)$. We symmetrize the measures according to [119] using

$$D_{sym}(p, \tilde{p}) = \frac{1}{\frac{1}{D(p,\tilde{p})} + \frac{1}{D(\tilde{p},p)}}. \tag{3.5}$$

Symmetric measures are termed $D1_{sym}$ and $D2_{sym}$. Performance of the symmetric and asymmetric measures is described in Section 3.4.3.

As a final step, the six computed KLDs, together with $\bar{\sigma}_{active}$ and $\bar{\sigma}_{inactive}$, are mapped to $\widehat{MOS}$. We experiment with several different candidate mapping functions: linear, multivariate polynomial and support vector regression (SVR). Simulation results showed that a radial basis SVR, with parameters optimized via linear search, provides lower estimation error. The results to follow are all based on using SVR. The reader is referred to [96] for a more comprehensive SVR review.

3.4.2 Algorithm Design Considerations

The KLD measure $D1$ described in (3.2) requires that both the reference GMM and the online-estimated model have the same number of Gaussian components. A larger number of components may hamper online parameter estimation. It is observed that speech databases used for subjective listening-quality assessment contain files that are on average 7 seconds long with an activity ratio between 60% and 85%. Hence, GMMs with 6 and 2 components are chosen for active and inactive models, respectively. This choice results in a training ratio (ratio between number of frames in the test signal and number of parameters estimated during training) of approximately 10. For the KLD measure $D2$ described in (3.4), we experiment with reference models with $6 \leq M \leq 16$ for active frames and $2 \leq M \leq 6$ for inactive frames. Superior performance is attained with 10 and 4 components, respectively. Moreover, we allow the number of GMM components for the test signal to vary such that the training ratio is kept above 10. It is observed that for most signals on our test databases (described in Section 3.4.3) the chosen number of components is $\tilde{M} = 6$ and $\tilde{M} = 2$, for active and inactive frames, respectively.

3.4.3 Experimental Results

In this section, experimental results are presented. Section 3.4.3.1 describes the databases used for training and testing of the proposed algorithm, Section 3.4.3.2 presents the experimental results, and Section 3.4.3.3 addresses multi-component quality estimation.

3.4.3.1 Database Description

The NOIZEUS database, described in Section 3.3.2, is used to design the reference GMMs. To train the MOS mapping function, a proprietary subjectively scored database is used. The database is comprised of speech corrupted by car and street noise at SNR=15 dB and office noise at SNR=20 dB and processed by the SMV speech codec; a total of 960 speech files are available. Three datasets not used in training (i.e., unseen) are used for testing. The first dataset (DS1) is comprised of speech corrupted by four noise sources (babble, car, street and hoth) at three SNR levels (0 dB, 10 dB, and 20 dB). The second (DS2) has noisy speech files (babble, street, car) at three SNR levels (0 dB, 10 dB, and 20 dB) processed by two noise suppression algorithms (SMV and Adobe Audition® with its "reduction level" parameter set to 75%). The third (DS3) is comprised of noisy speech signals (car, hoth, babble at 10 dB and 20 dB) processed by three speech codecs (G.711, G.729, and AMR) with packet loss concealment (PLC) capabilities. Random and bursty losses are simulated at 2% and 4%. A silence insertion concealment scheme is also present. Dataset DS3 is used to test the robustness of the proposed algorithm to alternate (unseen) methods of speech enhancement, in particular, packet loss concealment. The combined three test datasets consist of 1080 speech files covering 135 degradation conditions.

3.4.3.2 Test Results

Table 3.3 presents "per-condition" correlation (R) and root-mean-square error ($RMSE$) between MOS-LQS and P.563 MOS-LQO, for the three unseen datasets. Results are obtained after 3^{rd} order monotonic polynomial regression, as recommended in [50]. The table also reports the percentage improvement, relative to P.563, attained by the proposed algorithm for the four KLD measures described in Section 3.4.1.4. The columns labeled "$\%R$" and "$\%RMSE$" list the percentage increase in R and percentage reduction in $RMSE$, respectively. Note that "$\%R$" differs from the "R-improvement" measure described in (2.12). For measure $D2_{sym}$, R and $RMSE$ values are also shown to ease comparison. As can be seen, the proposed algorithm outperforms P.563 on all three datasets; as much as 60% increase in R and 37% decrease in

Table 3.3: Performance of P.563 and the proposed algorithm on three unseen datasets

Unseen Dataset	P.563		$D1$		$D1_{sym}$	
	R	$RMSE$	$\%R$	$\%RMSE$	$\%R$	$\%RMSE$
DS1	0.838	0.355	8.1	22.6	6.6	17.8
DS2	0.631	0.492	31.1	27.6	32.6	29.5
DS3	0.527	0.293	38.2	19.4	52.1	29.6
Average	–	–	**34.6**	**23.5**	**42.4**	**29.5**

Unseen Dataset	$D2$		$D2_{sym}$			
	$\%R$	$\%RMSE$	R	$\%R$	$RMSE$	$\%RMSE$
DS1	8.7	24.3	0.926	10.4	0.246	30.6
DS2	35.1	32.6	0.865	37.0	0.318	35.2
DS3	49.5	27.6	0.846	60.5	0.183	37.2
Average	**42.3**	**30.1**	–	**48.8**	–	**36.2**

$RMSE$ can be attained. The plot in Fig. 3.5 depicts MOS-LQO versus MOS-LQS for the proposed $D2_{sym}$ measure; each data point represents one of the 135 degradation conditions available in the combined test dataset.

Furthermore, it is observed that for datasets DS1 and DS2, similar performance is attained for asymmetric and symmetric measures. This is due to the fact that when p and $\tilde{p}$ are similar (i.e., test signal is "consistent" with one of the reference models, as expected for DS1 and DS2) the KLD takes on small values and $D(p, \tilde{p}) \approx D(\tilde{p}, p) \approx D_{sym}(\tilde{p}, p)$. On the other hand, when a test signal is not as consistent with the reference model (e.g., noisy speech processed by a PLC algorithm, as in DS3) the KLD takes on larger values and $D(p, \tilde{p}) \neq D(\tilde{p}, p)$. In this case, the symmetric measure performs better. Another example of this behavior can be observed with unseen test signals corrupted by speech-correlated noise (MNRU); measure $D2$ results

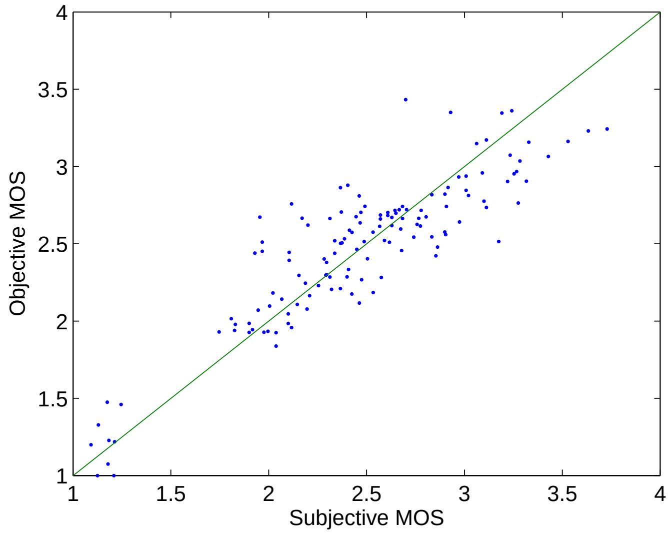


Figure 3.5: Per-condition MOS-LQO versus MOS-LQS for the combined test datasets using the proposed $D2_{sym}$ measure.

in $R = 0.782$ and $RMSE = 0.685$ while $D2_{sym}$ in $R = 0.955$ and $RMSE = 0.326$. For comparison purposes, P.563 achieves $R = 0.9142$ and $RMSE = 0.443$.

3.4.3.3 Component Quality Estimation

Due to the modular architecture of the proposed algorithm, a simple extension can be implemented to allow for single-ended measurement of BCK-LQS and SIG-LQS. In particular, two new SVR mapping functions are obtained. To estimate signal distortion, a 4-dimensional SVR is devised to map the KLDs computed from active frames (relative to the three reference models) and $\bar{\sigma}_{active}$ into SIG-LQO. To estimate background intrusiveness, a 5-dimensional SVR is designed to map the KLDs computed from inactive frames, $\bar{\sigma}_{inactive}$, and an estimated SNR to BCK-LQO. Here, we use the SNR estimated by the AMR VAD algorithm.

Since only the NOIZEUS database contains subjective SIG and BCK scores, 10-fold cross validation is used to measure the performance of the proposed scheme. The NOIZEUS database is randomly divided into 10 data sets of almost equal size. Training and testing is performed in 10 trials, where, in each trial, one of the data sets serves as a test set and the remaining 9 are combined to serve as a training set. Each data set serves as a test set only once. The ten resulting R's and $RMSE$'s are averaged to obtain the cross-validation performance figures. The proposed single-ended algorithm is shown to attain an average $R = 0.80$ and $RMSE = 0.33$ for SIG-LQO, and $R = 0.74$ and $RMSE = 0.39$ for BCK-LQO.

3.4.4 Algorithm Processing Time

Processing time is also an important figure of merit for gauging algorithm performance. We use the ANSI-C reference implementation of P.563. With the exception of the VAD algorithm (taken from the ANSI-C reference implementation of the AMR codec), the remainder of the proposed algorithm is implemented using Matlab version 7.2 Release 2006a. Simulations are run on a PC with a 2.8 GHz Pentium 4 processor and 2 GB of RAM. Here, processing time is defined as the time it takes to process ten speech files randomly selected from the three unseen test sets. The ten files combined have a total length of 57.77 seconds. For P.563, a processing time of 13.75 seconds is attained. The proposed algorithm (using $D2_{sym}$) has a processing time of 9.04 seconds, an approximate 35% reduction. A slight decrease in processing time of 0.15 seconds can be attained by using $D1_{sym}$. Note that a complete C implementation of the proposed algorithm would further increase the speedup.

Table 3.4 describes the percentage of the total processing time used by each module in the proposed algorithm. As can be seen, the computational complexity of the

Table 3.4: Algorithm processing times

Processing Module	Time (s)	%
Level normalization & IRS	1.30	14.4
PLP calculation	0.91	10.1
Cepstral deviation calculation	0.01	0.1
Voice activity detection	5.90	65.3
GMM parameter estimation (EM)	0.68	7.5
KLD calculation & MOS mapping	0.24	2.6
Total	**9.04**	**100**

proposed algorithm is mainly attributable to voice activity detection and level normalization and IRS filtering. A more efficient VAD algorithm and implementation would further decrease algorithm processing time. Further improvements may also be attained if the algorithm is employed in a codec-integrated manner, as proposed in Section 4.6. Moreover, experiments also show that only a slight decrease in performance is attained if level normalization and IRS filtering are not performed; this can result in a 44% reduction in processing time relative to P.563.

3.5 Conclusions

In this chapter, two configurations have been proposed for quality measurement of noise suppressed speech. The first, a network-distributed configuration, subsumes both double- and single-ended measurement paradigms. The results demonstrate that, besides offering the conventional functionality of measuring the quality of systems that degrade speech, the algorithm is capable of also measuring the quality of the transmission system per se and of characterizing the performance of the noise suppression algorithm. In this role, the proposed algorithm performs better than P.563 and provides a functionality not available with PESQ. The second, a single-ended configuration, proposes the use of Kullback-Leibler distances, computed between online- and offline-obtained statistical models of speech behaviour, as indicators of speech quality. Besides offering the conventional function of measuring the overall quality of a noise suppressed speech signal, both proposed configurations also allow for the estimation of quality dimensions labeled "signal distortion" and "background intrusiveness," as described in ITU-T P.835 [19].

Chapter 4

Hybrid Signal-and-Link-Parametric Quality Measurement for VoIP Communications

4.1 Preamble

This chapter is compiled from material extracted from a manuscript published in the IEEE Transactions of Audio, Speech, and Language Processing [84] and a manuscript to appear in the Elsevier Speech Communications Journal [83]. Earlier versions of this work appeared in the Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing [53].

4.2 Introduction

For VoIP communications, it is known that current signal-based measurement algorithms produce large "per-call" (also termed "per-sample") quality estimation errors and error variance [37–39]. Large estimation errors limit the usability of signal based methods for online quality monitoring and control purposes [70]. Link parametric methods, on the other hand, provide low per-call quality estimation errors and have been widely deployed in VoIP communication services. The major disadvantage of link parametric measurement is that signal distortions that are not captured by connection parameters are *not* accounted for. With burgeoning wireless-VoIP communications, representative signal distortions include varying levels and types of background noise, artifacts introduced by noise suppression algorithms, as well as temporal clipping artifacts resultant from VAD errors. Since wireless-VoIP communications are expected to become ubiquitous in the near future [4, 5], it is important to investigate the effects of such distortions on the performance of existing state-of-the-art quality measurement algorithms.

In this chapter, several wireless-VoIP degradation conditions are simulated and an in-depth statistical analysis is performed to assess the limitations of standard signal-based and link parametric algorithms. Focus is placed on the effects of modern wireless-VoIP degradations on *listening quality*, hence, factors such as jitter and delay, which affect conversational quality [120], are not considered. With the obtained insights, a single-ended hybrid signal-and-link-parametric speech quality measurement algorithm is proposed. The remainder of this chapter is organized as follows. Section 4.3 describes experiments which motivate the need for hybrid quality measurement. Section 4.4 introduces the architecture of the proposed algorithm, Section 4.5

presents experimental results, and Section 4.6 addresses the processing time of the proposed scheme. Lastly, conclusions are presented in Section 4.7.

4.3 Motivation for Hybrid Measurement

In this section, experiments are described which motivate the need for single-ended hybrid signal-and-link-parametric quality measurement. Experiments are carried out with a subjectively scored database, as described in Section 4.3.1. The sensitivity of P.563 to different VoIP network parameters is investigated in Section 4.3.2. Limitations of pure link parametric approaches, here characterized by the performance of an extended E-model implementation, are discussed in Section 4.3.3. Lastly, Section 4.3.4 reports P.563 and E-model performance figures for wireless-VoIP degradations; for comparison purposes, PESQ performance is also reported.

4.3.1 Database Description

A bilingual (English and French) subjectively scored speech database is used in our experiments. The speech database contains a wide range of typically-encountered VoIP scenarios. In particular, it comprises speech processed by G.711, G.729 and Adaptive Multi-Rate (AMR) codecs, with the latter operating at full rate (12.2 kbps). The packet loss concealment (PLC) algorithm used for G.711 is described in [121]; G.729 and AMR have their own built-in PLCs. Speech signals processed by G.711 with a simple silence insertion PLC scheme are also included. Packet durations of 10, 20, and 30 ms are used, except for AMR where only 20 ms packets are available. Random and bursty losses are simulated at 1, 2, 4, 7, and 10% with the ITU-T

G.191 software package [122]; the Bellcore model is used for bursty losses. Losses are applied to speech packets, thus simulating a transmission network with voice activity detection (VAD).

To investigate the limitations of pure link parametric measurement methods, several signal-based distortions are generated in combination with codec distortions (with and without packet losses). Signal-based distortions include temporal clippings, acoustic background noise, and noise suppression artifacts. In order to maintain the simplifying E-model assumption that impairments are additive in the perceptual domain, low levels of noise and packet loss rates are used. Temporal clipping distortions are either manually generated by replacing the beginning of a talkspurt with a copy of the noise floor, or simulated by forcing VAD false negatives in the G.711 and G.729 codecs. Acoustic noise distortions are generated by corrupting clean speech with three noise types (hoth, babble, and car) at two signal-to-noise ratio (SNR) levels (10 and 20 dB). Noisy speech is then processed by the three aforementioned speech codecs (singly or in tandem) with and without packet losses. In the former scenario, random and bursty packet losses are simulated at 2% and 4%.

Noise suppression artifacts in combination with codec distortion are used to further simulate impairments introduced by wireless-VoIP connections. Here, two noise suppression algorithms are tested. The first is the spectral subtraction algorithm available in the Adobe Audition software; a suppression factor of 75% is used. The second is the state-of-the-art noise suppression algorithm available as a pre-processing module in the SMV codec. Speech is corrupted by four noise types (hoth, car, street, and babble) at three SNR levels (0 dB, 10 dB, and 20 dB). Noisy speech is processed by the noise suppression algorithms and the noise-suppressed signal is input to the

G.711, G.729 or AMR speech codec.

The raw speech files were recorded in an anechoic chamber by four native Canadian French talkers and four native English talkers (half male and half female). Reference speech signals were filtered using the modified intermediate reference system (MIRS) send filter according to ITU-T Recommendation P.830 Annex D [123]. Degraded speech signals were further filtered using the MIRS receive filter. In both instances, speech signals were level adjusted to -26 dBov (dB overload) and stored with 8 kHz sampling rate and 16-bit precision. Similar to the ITU-T Supp. 23 dataset [106], each speech file comprises two sentences separated by an approximately 650 ms pause.

The subjective MOS test was conducted in 2006 following the requirements defined in [13, 123]. Sixty listeners (native in each language; roughly half male and half female) participated in each listening quality test. The headphones used were Beyerdynamic DT 770 and the ambient noise level in the listening room was kept at around 27-28 dBA. A total of 300 degradation conditions are available per language. Of the 300 impairment conditions, 146 are due to packet losses, 21 to temporal clipping, 31 to acoustic noise *and* codec distortion, 54 to acoustic noise, codec distortion *and* packet losses, and 48 are due to noise suppression *and* codec distortion. The tests described in Section 4.3.2 make use of the 146 packet loss degradation conditions (total of 1168 speech files), the tests in Section 4.3.3 make use of the 85 noisy speech conditions (total of 672 speech files), and the tests in Section 4.3.4 make use of the 85 aforementioned noisy conditions in addition to the 48 noise-suppressed conditions, totalling 1056 speech files.

4.3.2 Limitations of Pure Signal-Based Measurement

In this section, statistical analysis is used to assess the relationship between P.563 performance and four connection parameters: codec-PLC type, packet size, packet loss pattern (random or bursty), and packet loss rate. For real-time quality monitoring and control applications, objective quality measures are required to attain low per-call estimation errors. Hence, we use per-call MOS residual as the performance criterion. MOS residual is given by MOS-LQO minus MOS-LQS (or MOS-LQE minus MOS-LQS) and is abbreviated as "LQO-LQS" (or "LQE-LQS") in Figures 4.1-4.3. Factorial analysis of variance suggests that two parameters have significant main effects on P.563 accuracy, as described in Section 4.3.2.1. Two significant two-way interaction effects on P.563 accuracy are described in Section 4.3.2.2.

4.3.2.1 Main Effects

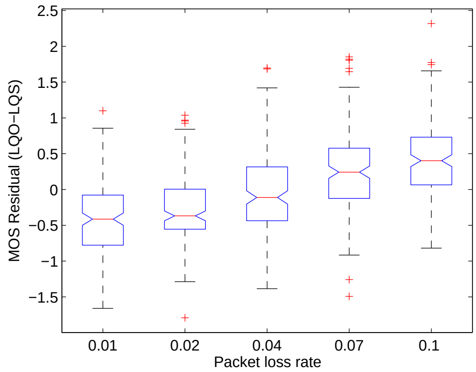
Factorial analysis of variance suggests that codec-PLC type and packet loss rate incur significant main effects on P.563 accuracy ($p < 0.0001$); the box and whisker plots depicted in Fig. 4.1 (a) and (b) assist in illustrating this behavior, respectively. The boxes have lines at the lower quartile, median, and upper quartile values; the whiskers extend to 1.5 times the interquartile range. Outliers (data with values beyond the ends of the whiskers) are represented by the symbol "+". The plots are computed using the Matlab function "boxplot" and the vertical width of the notches that cut into the boxes at the median line indicates the variability of the *median* between samples. When the notches of two boxes do not overlap, their medians are significantly different at the 95% confidence level [124]. From Fig. 4.1 (a), it can be seen that P.563 performance is strongly dependent on packet loss rates. P.563

underestimates MOS-LQS for low loss rates and overestimates MOS-LQS for higher loss rates; MOS residuals greater than 2 MOS points are obtained at a 10% loss rate.

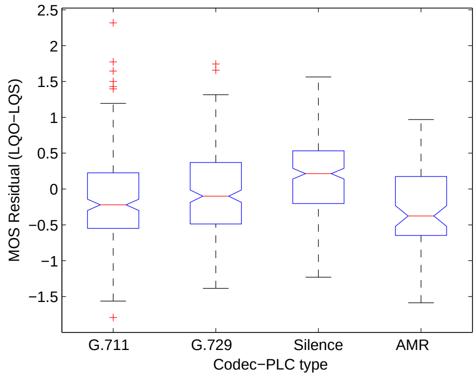
Fig. 4.1 (b) suggests that P.563 attains high per-call estimation errors, in particular for the G.711 PLC scheme. According to [50], P.563 has only been validated for PLC schemes in CELP (codebook-excited linear prediction) codecs (e.g., G.729); this can explain the poor performance obtained for G.711. Nonetheless, for the G.729 codec, P.563 attains residual errors that can be greater than 1.5 MOS point; on a five-point MOS scale, this can be the difference between having acceptable and unacceptable quality [39]. Moreover, the smallest median MOS residual occurs with the simple G.711 silence insertion loss concealment scheme; this can be explained by the fact that P.563 is equipped with a temporal clipping detection module. As will be shown in Section 4.3.2.2, however, this does not hold true for high packet loss rates. For comparison purposes, Fig. 4.1 (c) depicts the *non-significant* effects of speech codec-PLC type on extended E-model (described in Section 4.3.3) performance. The substantially smaller residual errors validate the accuracy of the extended E-model implementation and corroborate the popularity of link parametric measurement for VoIP online quality monitoring.

4.3.2.2 Two-way Interactions

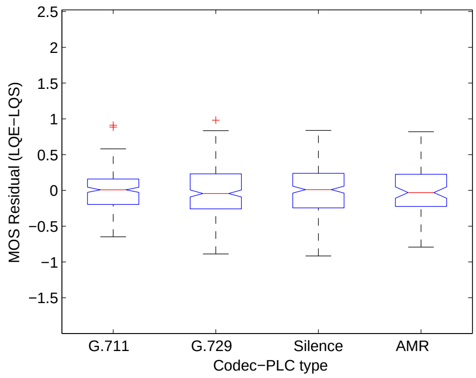
Statistical analysis has suggested two significant interaction effects on P.563 performance: codec-PLC type and packet loss rate ($p < 0.007$), and loss pattern and packet loss rate ($p < 0.003$). Box and whisker plots depicted in Fig. 4.2 (a) and (b) help illustrate this behavior, respectively. From Fig. 4.2 (a), it can be seen that P.563 underestimates MOS-LQS for low packet loss rates for both G.711 and G.729 codecs.



(a)

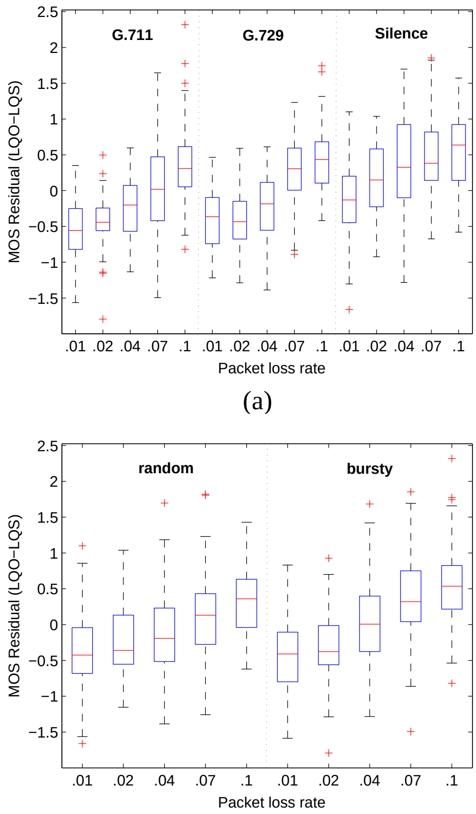


(b)



(c)

Figure 4.1: Significant one-way effects of (a) packet loss rates and (b) codec-PLC type on P.563 accuracy. For comparison purposes, (c) depicts non-significant effects of codec-PLC type on extended E-model accuracy.



(a)

(b)

Figure 4.2: Significant two-way interactions of (a) codec-PLC type and loss rate, and (b) loss rate and loss pattern on P.563 accuracy.

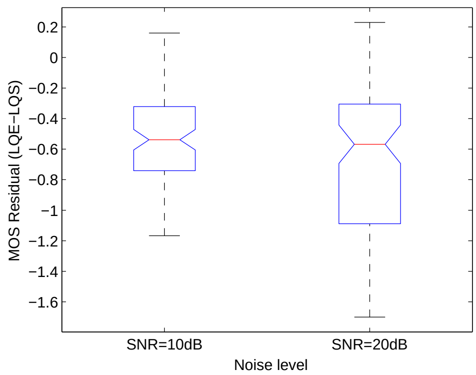
The simple silence insertion scheme attains median residual values closer to zero, except for high packet loss rates (10%) where it attains the highest residual median value. The largest residual errors (outliers) occur for the G.711 codec under a 10% packet loss rate.

Moreover, Fig. 4.2 (b) suggests that P.563 accuracy varies less for random losses than for bursty losses. For low packet loss rates, median residual MOS values are similar for both random and bursty packet losses. For higher loss rates, bursty losses attain median residual MOS values almost one-quarter of a MOS point higher than random losses. Relative to link parametric measurement, P.563 is shown to be more sensitive to connection parameters and to attain higher per-call estimation errors.

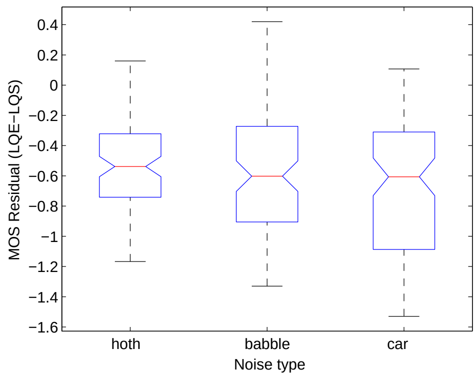
4.3.3 Limitations of Pure Link Parametric Measurement

Parameters used in the E-model represent terminal, network, and environmental quality factors that are assumed to be known *a priori*. Extended E-model implementations propose to estimate parameters (e.g., SNR) in real-time [73]. In this experiment, an extended E-model implementation is used. Non-tabulated equipment impairment factors are obtained from subjectively scored speech data (in accordance with [65]) and the noise level is computed using the clean reference speech signals. Note that link parametric measurement is favored with this unrealistic assumption that true noise level information is available online. Commonly, only estimated noise levels are available, as in the experiment described in Section 4.5. Here, statistical analysis is used to investigate the effects of noise level, noise type, and noise *and* packet loss on extended E-model measurement performance. The analysis suggests significant main effects of noise level ($p < 0.006$) and noise type ($p < 0.04$); the box and whisker plots depicted in Fig. 4.3 (a) and (b) help illustrate this behavior, respectively.

From the plots, it can be observed that the extended E-model underestimates MOS-LQS and has a higher residual MOS variance for lower noise levels (SNR= 20 dB) and for babble and car noise. On the other hand, we observe that noise level does *not* show significant effects on P.563 accuracy. P.563 is equipped with a "noise analysis" module which not only estimates the SNR, but also takes into account other spectrum-related measures such as high frequency (2500-3500 Hz) spectral flatness. It is observed, however, that P.563 performance is lower for babble and car noise, both of which have "low-pass" characteristics.



(a)



(b)

Figure 4.3: Significant one-way effects of (a) noise level and (b) noise type on extended E-model accuracy.

4.3.4 Performance Figures

In this section, we investigate the accuracy of the E-model and P.563 algorithms under the wireless-VoIP degradation conditions described in Section 4.3.1. For comparison, PESQ results are also reported where original clean speech files are used as reference. As suggested in [125], the true noise level is used (assumed to be known *a priori*) in E-model calculations and subjective tests are used to obtain impairment factors not described in [62–64].

Moreover, to our knowledge, equipment impairment factor values for noise suppression algorithms are not available. In fact, artifacts introduced by such enhancement schemes are dependent on noise type and noise levels. In our experiments, an estimated SNR (post enhancement) is used in the computation of MOS-LQE for noise-suppressed speech. The SNR is estimated with the P.563 "noise analysis" module described in [50]. In a controlled experiment, the estimated SNR is shown to be highly correlated with the true SNR (correlation close to unity). It is important to emphasize, however, that while using estimated (or measured) SNR is convenient for quantifying noise artifacts that remain after enhancement, noise suppression artifacts that arise during speech activity are *not* accounted for; this is a major shortcoming of parameter-based quality measurement methods.

Pearson correlation (R) and root-mean-square error ($RMSE$) are used as performance figures. Results in Table 4.1 are reported on a per-condition basis where condition-averaged MOS-LQS and condition-averaged MOS-LQO (LQE) are used to estimate R and $RMSE$. For comparison, performance figures are reported before and after 3^{rd} order monotonic polynomial regression. Moreover, as suggested in [40], PESQ performance is reported before and after the mapping described in [108]. The

Table 4.1: Per-condition performance of E-model, PESQ and P.563. Post-mapping performance is represented by R^* and $RMSE^*$.

	E-model				P.563			
	R	$RMSE$	R^*	$RMSE^*$	R	$RMSE$	R^*	$RMSE^*$
Noisy	0.716	0.715	0.722	0.278	0.622	0.479	0.643	0.307
Noise-suppressed	0.917	0.547	0.924	0.301	0.767	0.587	0.799	0.472
Overall	0.657	0.661	0.676	0.411	0.642	0.507	0.673	0.413

	PESQ			
	R	$RMSE$	R^*	$RMSE^*$
Noisy	0.641	0.315	0.633	0.465
Noise-suppressed	0.937	0.393	0.921	0.475
Overall	0.833	0.324	0.816	0.464

post-mapping performance figures are represented by R^* and $RMSE^*$ in Table 4.1.

As can be seen for noisy speech, E-model based measurement outperforms signal-based measurement. Recall, however, that in this experiment, E-model based measurement is favored as true noise information was used in E-model computations. It is also observed that P.563 performance is comparable to that of PESQ; this is an important result as P.563 does not make use of a clean reference signal. For noise-suppressed speech, however, P.563 performance is inferior to PESQ and E-model. Overall, performance figures are substantially lower than those reported for traditional telephony applications for all three standard algorithms (e.g., see [51, 126]).

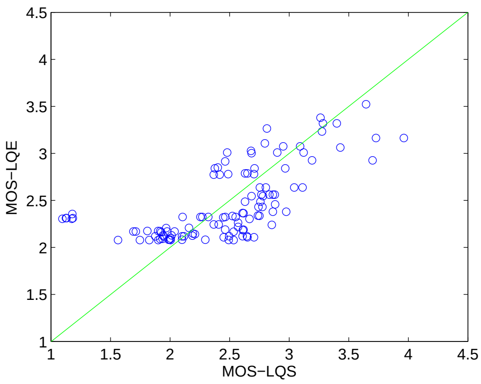
The plots in Fig. 4.4 (a)-(c) depict the overall per-condition MOS-LQO (LQE) *versus* MOS-LQS for the E-model, P.563, and PESQ, respectively. Plots (a) and (b) are after 3^{rd} order polynomial mapping and plot (c) depicts PESQ MOS-LQO before ("$\circ$") and after ("$\times$") the mapping described in [108]. As can be seen from the plots and from Table 4.1, PESQ performance decreases once the mapping is applied.

This suggests that an alternate mapping function needs to be investigated for modern degradation conditions such as those present in wireless-VoIP communications.

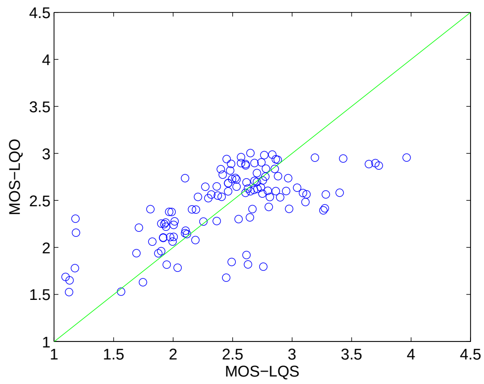
4.3.5 Discussion

The comparisons described above suggest that the performance of standard objective quality measurement algorithms is compromised for degradation conditions present in wireless-VoIP communications. To quantify the decrease in measurement accuracy, a calibration experiment is carried out with conventional subjectively scored VoIP speech data (English and French). Here, *clean* speech, as opposed to noise-corrupted or noise-suppressed speech, is processed by the four aforementioned codec-PLC schemes (G.711, G.711*, G.729, and AMR) under the same random and bursty packet loss conditions (2% and 4%). With such conventional VoIP impairment scenarios, standard algorithms are shown to perform reliably (e.g., see [51, 126]). In this calibration experiment, it is observed that MOS-LQE estimates attain an $RMSE$ that is 48% lower than that reported in Table 4.1. The MOS-LQO estimates (for both PESQ and P.563), in turn, attain an $RMSE$ value that is 35% lower.

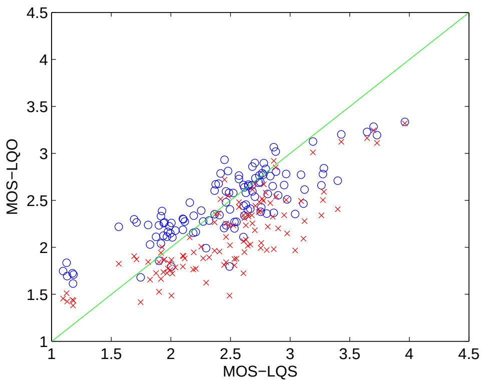
As observed, E-model performance is affected more severely by wireless-VoIP distortions. Such behavior is expected as the E-model is a parameter-based measurement method and, as such, overlooks signal-based distortions that are not captured by the link parameters. Hence, improved performance is expected from hybrid signal-and-link-parametric measurement schemes where signal-based distortions are estimated from the speech signal and used to improve parameter-based quality estimates. The architecture of the proposed hybrid measurement algorithm is described next.



(a)



(b)



(c)

Figure 4.4: Per-condition MOS-LQO *versus* MOS-LQS for the overall dataset after $3^r d$ order polynomial mapping for (a) the E-model and (b) P.563, and (c) PESQ before ("○") and after ("×") the mapping described in [108].

4.4 Architecture of the Proposed Hybrid Measurement Algorithm

The overall architecture of the proposed algorithm is depicted within the dotted lines in Fig. 4.5. Offline, E-model ratings and subjective listening tests are used to determine the base quality representative of several VoIP communications scenarios. As an embodiment of the proposed approach, we obtain base quality values for commonly used codec-PLC types with different packet sizes, under different packet loss patterns and packet loss rates. Base quality values are stored in a lookup table for fast online operation. Statistical models, in particular Gaussian mixture models, are designed using perceptual features extracted from speech signals processed by the various speech codecs operating under clean reference conditions. Reference GMM parameters, $\boldsymbol{\lambda}$ as defined in Section 2.3.3, are also stored in a lookup table for each codec. The speech codecs used in our experiments are those described in Section 4.3.1.

Online, IP header-extracted parameters are used to obtain the base quality and reference GMM parameters from lookup tables. Once packets are decoded and PLC is performed, the speech signal is level-normalized and filtered. Perceptual features are then extracted from the pre-processed test signal. The voice activity detector labels the feature vector of each frame as either active or inactive. The extracted features are compared to stored models of normative codec operation behavior via a simple consistency measure. Temporal discontinuity detection is used to detect temporal clippings, an impairment which occurs commonly in VoIP communications [101]. Lastly, a MOS-mapping module is used to map the base quality, computed consistency measures, noise spectrum tilt, and detected temporal discontinuities to a

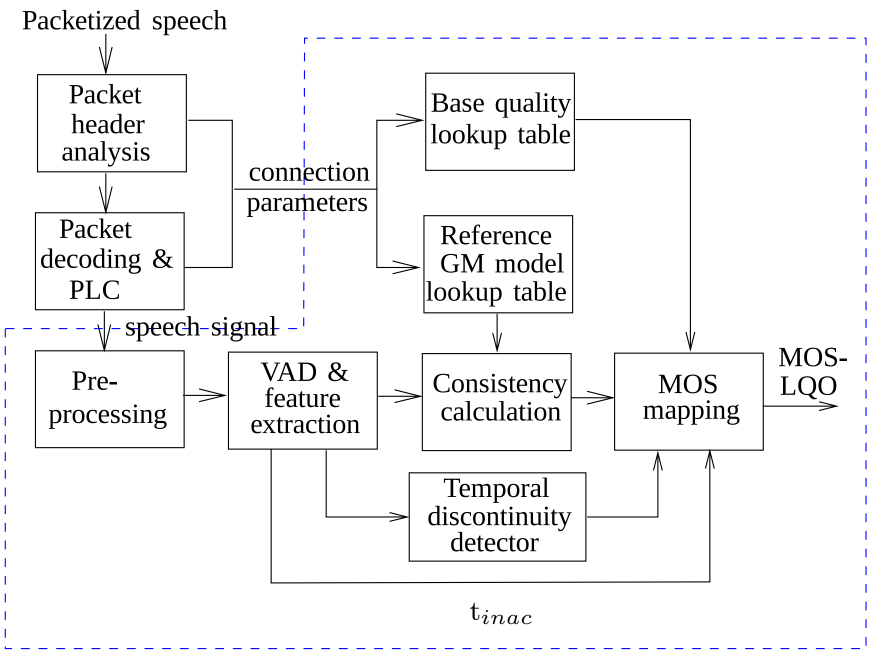


Figure 4.5: Architecture of the proposed hybrid signal-and-link-parametric quality measurement algorithm.

final MOS-LQO. A more detailed description of each signal-based processing block is provided in the remainder of this section.

4.4.1 Pre-processing, VAD and Feature Extraction

The pre-processing module performs level normalization and IRS filtering. The level of the speech signal is normalized to -26 dBov using the P.56 voltmeter [114] and the MIRS filter is applied to emulate the handsets used in listening tests. Voice activity detection is employed to label speech frames as active or inactive; the VAD from the G.729 codec [127] is used.

Fifth order perceptual linear prediction (PLP) cepstral coefficients $\mathbf{x} = \{x_i\}_{i=0}^{5}$ are extracted from the speech signal and serve as primary features. The zeroth cepstral coefficient x_0 is employed as an energy measure and differential PLP cepstral coefficients are used as a measure of signal spectral dynamics; in particular, delta and double-delta cepstral coefficients are used. Motivated by the results described in Section 4.3.3, noise-related features are also extracted. Pilot experiments are carried out with noise spectral flatness and noise spectrum tilt; the latter (henceforth referred to as t_{inac}) resulted in superior performance and is used throughout the remainder of this chapter. As in [116], t_{inac} is approximated by the 1^{st} order linear prediction coefficient averaged over inactive speech frames.

4.4.2 GMMs and Consistency Calculation

Reference models of normative codec behavior are designed for commonly used speech codecs. For active speech frames, GMMs are trained for PLP cepstral coefficients appended with delta and double-delta coefficients, i.e., $\mathbf{z}_{act,m} = [\mathbf{x}_m, \Delta\mathbf{x}_m, \Delta^2\mathbf{x}_m]$. For inactive speech frames, GMMs are obtained from PLP cepstral coefficients $\mathbf{x}_m$. Gaussian mixture models are given by (2.4) and are described in more detail in Section 2.3.3. For the sake of notation, a Gaussian mixture density will be represented in subsequent sections as $\boldsymbol{\lambda}$=$\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$, where $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i\}$. In our experiments, diagonal covariance matrix Gaussian components are used and two recursive (greedy) EM algorithm implementations are investigated. Greedy EM implementations estimate model parameters $\boldsymbol{\lambda}$ and the number of Gaussian components M simultaneously.

We experiment with a pruning approach which starts with a large number of components and uses a stochastic approximation recursive learning algorithm to prune irrelevant components [128], and a progressive approach which starts with a single component and adds components sequentially [129]. In order to guard against over-fitting, the largest admissible M is chosen such that the training ratio (ratio between number of scalar parameters that need to be estimated during training and the number of training samples available) is maintained above an empirically-set value of 100. In our experiments, it is observed that both implementations result in similar performance and $M = 32$ is chosen for active frames and $M = 4$ for inactive frames. Lastly, reference-model consistency measures are computed as per (2.5). The notation c_{act} and c_{inac} is used to denote consistency measures computed for active and inactive speech frames, respectively.

4.4.3 Temporal Discontinuity Detector

Temporal discontinuities (also known as "clippings") are a known source of quality degradation in VoIP systems [101]. Front-end, midspeech (short mutes), and back-end clippings may occur due to erroneous VAD decisions, erroneous line echo cancelation decisions, or simple silence insertion PLC schemes. A simple energy-thresholding scheme is proposed and temporal discontinuities are detected by evaluating abrupt changes in x_0.

From our experiments, abrupt stops (back-end clippings) can be accurately detected if

$$\mathcal{T}_i = \frac{x_{0,i+1}}{x_{0,i}} < 0.35.$$

Abrupt starts (front-end clippings) are detected if $\mathcal{T}_i > 2.02$. Experiments on our databases show that with this simple energy-thresholding scheme, approximately 98% of front-end clippings are correctly classified. On the other hand, approximately 10% of "normal" abrupt starts, such as those experienced with certain plosive consonants (e.g., /d/), are misclassified as clippings. To improve classification performance, more complex machine learning methods can be used, such as the one described in Section 2.3.4. Since abrupt starts have, intuitively, less significant impact on perceived speech quality than abrupt stops [49, 130], such classification errors are shown not to be detrimental to overall speech quality measurement. For the sake of reduced computational complexity, the simple energy-thresholding scheme is used in the experiments described in Section 4.5. Lastly, midspeech clippings are detected when an abrupt stop is followed by an abrupt start during speech activity. The mute length is estimated from the number of consecutive frames for which $\mathcal{T} \simeq 1$.

Previous studies suggest that clipping frequency of occurrence and midspeech clipping duration are two major factors affecting subjective quality [130]. We define clipping frequency of occurrence as the ratio of the number of detected discontinuities over active speech duration (clips/second); frequency of occurrence is computed for front- and back-end clippings (f_f and f_b, respectively). For midspeech clippings, subjective tests suggest that similar quality is attained for high occurrence of short mutes and low occurrence of long mutes [130]. As a consequence, clipping frequency of occurrence is computed for midspeech clippings of short duration (f_{mid-s}) and long duration (f_{mid-l}). Mutes between 10-70 ms are classified as short duration and mutes between 70-260 ms as long duration.

4.4.4 MOS Mapping

Machine learning tools are used to devise an accurate mapping between the base quality (MOS_0), computed consistency measures, noise spectrum tilt, and clipping frequency of occurrence to a final MOS-LQO. Here, a support vector regressor (SVR)

[96], trained on subjectively scored data, is used. The input to the MOS mapping module is the 8-dimensional feature vector consisting of

$$\mathbf{z} = [\mathrm{MOS}_0, c_{act}, c_{inac}, \mathrm{t}_{inac}, \mathrm{f}_f, \mathrm{f}_b, \mathrm{f}_{mid-s}, \mathrm{f}_{mid-l}].$$

A subset of the ITU-T Supplement 23 (experiment 3) database [106], along with material from three other proprietary databases, is used to train the MOS mapping module. The Supplement 23 subset includes speech processed by the G.729 codec (singly or in tandem conditions) with random and bursty losses at 3% and 5%. Clean and noisy conditions (street, hoth and vehicle noise at an SNR=20 dB) are included. Proprietary databases include temporal-clipped speech material, speech processed by the G.711 codec with 3% random packet losses, and noisy speech. The latter includes speech degraded by car and street noise (SNR=15 dB) and processed by the SMV codec, operating at full and half rate (8.5 kbps and 4 kbps, respectively), with 1% random losses. A total of 2672 speech samples are used to train the MOS mapping function.

4.5 Experiments

The proposed hybrid signal-and-link-parametric measurement algorithm is tested on a subset (1232 speech samples) of the corpus described in Section 4.3.1. The subset includes 154 impairment conditions covering temporal clipping, noise *and* codec distortion, noise *and* packet losses, and noise suppression *and* codec distortion. Hence, the test set covers distortions which are not captured by connection parameters, such as those present in modern wireless-VoIP communications. We emphasize that degradation conditions and speech files available in the test set are distinct from those in

Table 4.2: Per-condition performance of hybrid, pure link parametric (extended E-model) and pure signal-based (P.563) measurement. Results are reported before and after third-order polynomial regression.

	Hybrid		Link (Extended E-model)				Signal (P.563)			
	R	$RMSE$	R	$\%R\uparrow$	$RMSE$	$\%RMSE$	R	$\%R\uparrow$	$RMSE$	$\%RMSE$
before	0.813	0.329	0.696	38.5	0.665	50.5	0.701	37.5	0.494	33.4
after	0.821	0.301	0.712	37.8	0.352	14.5	0.720	36.1	0.407	26.0

the training set, thus are unseen to the proposed algorithm. Comparisons are carried out with P.563 and the extended E-model described in Section 4.3.4.

Correlation (R) and root-mean-square error ($RMSE$) are used as algorithm figures of merit. We report improvement in correlation incurred by the proposed scheme over pure signal-based or link parametric measurement by the so-called "R-improvement" measure described in (2.12). Improvement in $RMSE$ attained by using the proposed algorithm is reported by means of the conventional percentage reduction in root-mean-square error ($\%RMSE$). As recommended in [50], results in Table 4.2 are reported on a per-condition basis where condition-averaged MOS-LQS and condition-averaged MOS-LQO (or MOS-LQE) are used to estimate R and $RMSE$. Results are reported before and after third-order monotonic polynomial regression.

As can be seen, the proposed method improves on pure link parametric measurement by approximately 38% and 15% in terms of R and $RMSE$, respectively (post third-order mapping). Improvements of approximately 36% and 26% (R and $RMSE$ respectively) are attained relative to pure signal-based measurement. For comparison purposes, PESQ attains $R = 0.831$ and $RMSE = 0.458$ with the mapping described in [108]. Thus, the hybrid single-ended scheme offers somewhat lower $RMSE$ than the state-of-the-art double-ended standard algorithm. Recall, however, that the usage

Table 4.3: Per-call $RMSE$ of hybrid, pure link parametric and pure signal-based measurement.

	Hybrid	Link	Signal
$RMSE$	0.513	0.689	0.587
$\%RMSE$	–	25.6	12.6

of PESQ is not recommended for systems that include a noise suppression algorithm. Furthermore, statistical analysis shows that noise type and noise level have *insignificant* effects on the performance of the proposed hybrid scheme.

As mentioned previously, for online quality monitoring applications, per-call residual MOS error is an important performance metric. Per-call $RMSE$ is reported in Table 4.3 for the extended E-model, P.563, and the proposed hybrid scheme. As can be seen, the proposed algorithm attains reductions in per-call $RMSE$ of approximately 13% and 26% relative to pure signal-based and pure link parametric measurement, respectively. From Table 4.3, it can also be observed that under noisy and wireless-VoIP conditions, P.563 attains smaller per-call residual errors than the extended E-model, thus corroborating the fact that pure link parametric measurement is compromised for degradations not captured by connection parameters.

4.6 Algorithm Processing Time

As mentioned previously, pure link parametric measurement has gained widespread use due to its low algorithm processing time. As a consequence, it is important to measure the computational overhead incurred by the signal-based branch of the proposed algorithm. Here, processing time is defined as the time it takes to process ten speech files randomly selected from the test set described in Section 4.5. With

Table 4.4: Algorithm processing time for P.563 and the proposed hybrid scheme, with and without codec-integrated processing.

Algorithm	Proc. time (s)	%Time↓
P.563	26.25	–
Proposed	12.47	52.5
Proposed (codec-integrated)	3.06	88.3

the exception of the VAD algorithm (taken from the ANSI-C G.729 reference implementation), the proposed algorithm is implemented using Matlab version 7.2 Release 2006a. Simulations are run on a PC with a 2.8 GHz Pentium 4 processor and 2 GB of RAM.

The ten files combined have a total length of 62.57 seconds. Algorithm processing times for the Matlab implementation of the proposed algorithm and the ANSI-C reference implementation of P.563 are reported in Table 4.4. The column labeled "%Time↓" describes the percentage reduction in processing time obtained by the proposed scheme relative to P.563. As can be seen, a reduction in processing time of approximately 53% is attained; note that a complete C implementation of the proposed algorithm would further increase the speedup.

An in-depth analysis of the processing time of each operational module depicted in Fig. 4.5 shows that over 75% of the processing time is attributable to voice activity detection. To further reduce processing time, the proposed algorithm can take advantage of the fact that in most VoIP codec implementations, VAD decisions are transmitted by the encoder and are readily available at the decoder. Moreover, in the event of a lost packet, VAD decisions are predicted by the decoder based on previously received packets. Hence, the hybrid scheme can reuse these VAD decisions in lieu of recomputing them. To investigate the gains obtained with such "codec-integrated" processing, we use the Matlab implementation of the G.723.1 speech codec described in [131] where inactive frames are detected as “null frames” in the G.723.1 bitstream. Table 4.4 also exhibits the gains obtained with the codec-integrated hybrid quality measurement scheme. As can be seen, an overall reduction in processing time of approximately 88% can be attained relative to P.563.

4.7 Conclusions

In this chapter, the limitations of standard signal-based and link parametric quality measurement algorithms are reported for emerging wireless-VoIP communications. A hybrid signal-and-link-parametric quality measurement scheme is then proposed to overcome such limitations. Experiments described herein serve to demonstrate the gains obtained by combining the strengths of pure signal-based and pure link parametric measurement paradigms to devise a more comprehensive quality measurement scheme. The proposed hybrid methodology improves on pure link parametric approaches by measuring distortions that are not captured by connection parameters. In turn, lower per-call estimation errors are attained relative to pure signal-based measurement. The proposed scheme is shown to have modest computational overhead relative to pure link parametric measurement, and when operated in an integrated manner, can attain processing time that is 88% lower than the ITU-T P.563 algorithm. Moderate computational complexity, low per-call estimation errors and the ability to account for distortions not captured by connection parameters are valuable attributes for online VoIP quality monitoring and control.

Chapter 5

Quality Measurement for Hands-Free Speech Communications

5.1 Preamble

This chapter is compiled from material extracted from a manuscript submitted to the IEEE Transactions on Instrumentation and Measurement in 2008 [85] and a manuscript that appeared in the Proceedings of the 2008 International Workshop for Acoustic Echo and Noise Control [86]. An earlier version of this work appeared in the Proceedings of the 2007 Interspeech/Eurospeech Conference [57].

5.2 Introduction

When speech is produced in an enclosed environment, the acoustic signal follows multiple paths from source to receiver. Such reflections may arrive with delays ranging from a few milliseconds to a few seconds, depending on room geometry and sound absorption properties. Early reflections, on the order of few tens of milliseconds, modify the signal short-time spectrum causing a change in signal timbre; such an effect is termed spectral colouration [132, 133]. Delays greater than 50 milliseconds (termed late reflections), on the other hand, are perceived as distinct copies of the direct path signal and cause temporal colouration distortions. The exponential decay of late reflections results in temporal smearing, which in turn, decreases perceived speech quality and intelligibility.

Traditionally, the time-domain room impulse response (IR) or room geometry and wall absorption properties are used to measure room acoustical parameters. Offline measurement of room impulse responses, however, is a laborious task. In addition, the impulse response varies with acoustic source positioning, room temperature, as well as placement of room furnishings. As a consequence, room acoustical parameters obtained from room IR measurements are not feasible for real-time signal processing applications. To this end, blind signal-based measurement, where room acoustical parameters are obtained from the reverberant speech signal, has been the focus of more recent research. Special emphasis has been given to blind estimation of the so-called reverberation time (T_{60}) parameter (see Section 5.3.2).

In the past, a handful of blind T_{60} estimators have been proposed. In [134], the diffuse tail of the reverberation is modeled as exponentially damped Gaussian white noise. A maximum-likelihood (ML) estimate of the time constant of the decay is used to characterize T_{60}. With ML-based approaches, it is common to assume that the source signal stops abruptly and has long pauses between speech segments; such requirements are needed in order to attain reliable estimates. As expected, the performance of ML-based methods is compromised for noise-corrupted reverberant speech. Notwithstanding, the work described in [135] proposes a "generalized" ML procedure which loosens the aforementioned assumptions and allows for blind T_{60} estimation under noisy environments.

Alternately, the work described in [136] shows that reverberation corrupts the harmonic structure of voiced speech segments. Hence, a measure of pitch "strength" (or periodicity) is used to blindly estimate T_{60}. The estimator, however, is shown to be sensitive to speaker gender. Additionally, the kurtosis of linear prediction (LP) residuals is used in [137] for blind T_{60} characterization. The idea is that for clean voiced speech segments, LP residuals have strong peaks corresponding to glottal pulses. The peaks become smeared in time as reverberation increases, thus reducing the LP residual kurtosis to that of a Gaussian distribution. LP residual-based methods have also been successfully used in the past for noise and reverberation suppression [138–140].

In this chapter, the use of temporal dynamics information is investigated for blind measurement of room acoustical parameters. *Short-term* dynamics information is obtained from commonly used delta cepstral coefficients. Statistics computed from the zeroth-order delta cepstral sequence (Δ_0) are shown to provide useful cues for blind T_{60} estimation. Moreover, *long-term* dynamics information is obtained by means of spectral analysis of temporal envelopes of speech, a process commonly termed *modulation spectrum* processing. A reverberation-to-speech modulation energy ratio measure is proposed and used for blind measurement of several room acoustical parameters, including estimators of subjective perception of spectral colouration, reverberant tail effect, and overall speech quality. Experiments described herein show that the proposed estimators outperform a baseline system in scenarios involving reverberant speech with *and* without the presence of acoustic background noise.

The remainder of this chapter is organized as follows. Section 5.3 describes models, characterization and simulation of room reverberation. Section 5.4 provides motivation and description of the proposed measures; experimental results are presented in Section 5.5. Objective assessment of perceived reverberation effects is discussed in Section 5.6 and conclusions are given in Section 5.7.

5.3 Room Reverberation

In this section, models of room reverberation are presented. Parameters commonly used to characterize reverberation are presented, as well as methods to generate reverberant speech.

5.3.1 Models of Room Reverberation

Conventionally, the propagation from source to microphone in a reverberant enclosure is modeled as a linear filtering process. The reverberant signal $s(n)$ is modeled as a convolution of the anechoic source speech signal $v(n)$ with the room IR $r(n)$

$$s(n) = v(n) * r(n). \tag{5.1}$$

If additive background noise $N(n)$ is present, (5.1) becomes

$$s(n) = v(n) * r(n) + N(n). \tag{5.2}$$

It is known that under the diffuse sound field assumption, the ensemble average of the squared room impulse response decays exponentially with time [141]

$$\langle r^2(n) \rangle = A \exp(-kn). \tag{5.3}$$

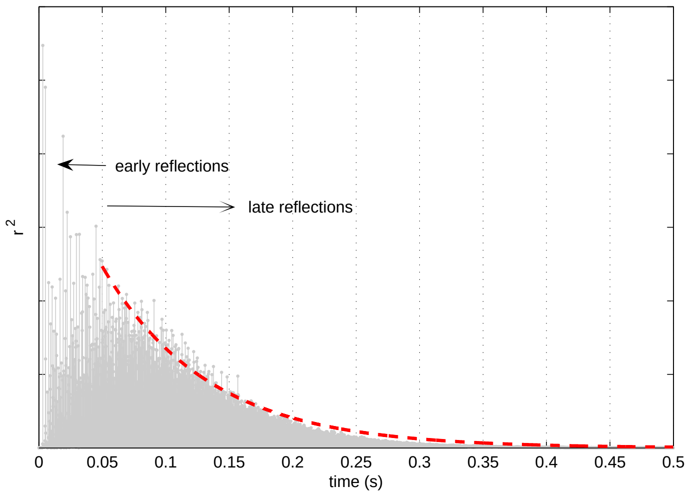


Figure 5.1: Exponential decay of the late reflections of a room with $T_{60} = 0.5$ s.

The angled brackets $\langle \cdot \rangle$ denote the ensemble average, A is a gain term, and k is the damping factor given by [141]

$$k = \log 10^6/(F_s \times T_{60}), \tag{5.4}$$

where F_s is the sampling frequency and T_{60} is the so-called reverberation time, as described in Section 5.3.2. The plot in Fig. 5.1 illustrates the exponential decay of a room impulse response generated via the image method [142] with $T_{60} = 0.5$ s and $F_s = 8$ kHz. The dashed curve in the figure illustrates the exponential decay given by (5.3) with $A = 0.0045$.

5.3.2 Characterization of Room Reverberation

Reverberation time (T_{60}) is the parameter most widely used to characterize room acoustics. By definition, it is the time required for the sound energy to decay by 60 dB after the sound source has been turned off [143]. Commonly, the so-called Schroeder integral is used to measure T_{60} from the room IR [144]. Other parameters that characterize room acoustics and are obtained from the room IR include early decay time (interval required for the energy to decay by 10 dB), speech clarity index (energy ratio between 50 ms early reflections and the remaining late reflections) [145], and direct-to-reverberation energy ratio (DRR). DRR, expressed in decibel, is the energy ratio between the direct sound and the room reverberation and is given by

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_d} r^2(n)}{\sum_{n=n_d+1}^{\infty} r^2(n)}\right) \text{[dB]}, \tag{5.5}$$

where $n_d F_s$ is the direct sound arrival time.

Moreover, the spectral content of the room IR can provide information regarding spectral colouration. In [146, 147], the second-order moment of the room frequency response is proposed as a measure of spectral colouration. Additionally, subjective listening tests may be used to characterize the perceived quality of speech signals produced in reverberant enclosures. In [148], subjective listening tests are used to characterize the perception of timbre. Recently, listening tests have been used to characterize subjective perception of colouration, reverberation decay tail effects, and overall quality for reverberant *and* reverberation-suppressed speech [22]. The test follows the guidelines described in ITU-T Recommendation P.835 [19].

5.3.3 Simulation of Reverberant Speech

Two tools are used to generate reverberant speech: SIREAC (SImulation of REal ACoustics) [149] and the ITU-T software package described in Recommendation G.191 [122]. Anechoic speech from eight speakers (half male, half female) are used throughout our experiments. A total of 256 utterances (averaging six seconds each)

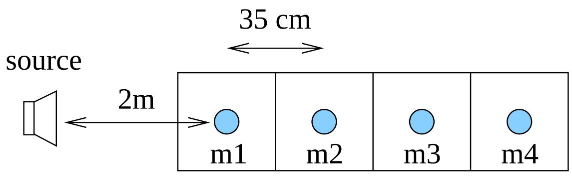


Figure 5.2: Microphone array setup at the Bell Labs varechoic chamber.

are spoken per speaker; half of the utterances are in English and the other half in French. Speech samples are each composed of two sentences separated by an approximately 800 ms pause; all signals are stored with an 8 kHz sampling rate and 16-bit precision. SIREAC is used to artificially generate reverberant speech with T_{60} values between 0.2-1 s (0.1 s increments), 1.5 s, and 2 s. The level of the reverberant speech signal is normalized to -26 dBov (dB overload) using the ITU-T P.56 voltmeter [114].

The ITU-T G.191 tool is used to convolve room impulse responses collected from real environments with the anechoic speech signals. The real room impulse responses are stored with an 8 kHz sampling rate and include those collected with a four-channel linear microphone array (as depicted in Fig. 5.2) at the Bell Labs varechoic chamber[1] with 100%, 43% and 0% panels open [150], and those collected with a single microphone in a large cafeteria, a medium-sized meeting room, a small lavatory, and a medium-sized office [151]. As with the simulated data, reverberant speech signals are normalized to -26 dBov. Table 5.1 reports parameters T_{60} and DRR, computed from the room impulse responses, for the aforementioned environments. In the table, varechoic chamber data is represented as "VC-%-mi" where "%" represents the percentage of open reflective panels and "mi" the microphone number in the microphone array (see Fig. 5.2).

[1]The Bell Labs varechoic chamber is a rectangular room with 368 independently actuated surfaces in the walls, ceiling, and floor. T_{60} is controlled by the percentage of open panels.

Table 5.1: Room acoustical parameters for real room impulse responses

Room	$T_{60}(s)$	DRR (dB)
VC-100%-m1	0.3	1
VC-100%-m2	0.3	1
VC-100%-m3	0.3	-1
VC-100%-m4	0.3	-1
VC-43%-m1	0.5	0
VC-43%-m2	0.5	-3
VC-43%-m3	0.5	-2
VC-43%-m4	0.5	-5
VC-0%-m1	0.9	-5
VC-0%-m2	0.9	-7
VC-0%-m3	0.9	-7
VC-0%-m4	0.9	-9
Office	0.6	-4
Meeting	0.9	-7
Lavatory	1.3	-9
Cafeteria	1.5	-14

5.4 Temporal Dynamics and Proposed Estimators

In this section, a description of the features used to capture short- and long-term temporal dynamics is given; the proposed T_{60} and DRR estimators are also described.

5.4.1 Short-Term Temporal Dynamics

Short-term energy dynamics information is used for blind measurement of T_{60}. In this chapter, the zeroth order mel-frequency cepstral coefficient (MFCC) is proposed as a measure of short-term log-spectral energy and the zeroth order delta MFCC as a measure of log-energy rate of change [91]. Mel-frequency cepstral coefficients (MFCC) can be obtained in a manner similar to conventional cepstra [91] with the exception that frequency bands are warped to a mel scale [152]. MFCC are chosen as they are widely used by the speech and speaker recognition communities, hence blind T_{60} estimators can be developed to improve recognition performance whilst requiring negligible computational overhead.

Let $c_{0,m}$ denote the zeroth order MFCC for frame m and $\mathbf{c}_0$ the cepstral sequence for a given speech signal. Analogously, let $\Delta c_{0,m}$ denote the per-frame zeroth order delta MFCC, computed according to (2.6), and $\Delta\mathbf{c}_0$ the delta sequence. Fig. 5.3 (a) depicts, from top to bottom, the waveform, $\mathbf{c}_0$, and $\Delta\mathbf{c}_0$, for a clean speech signal, respectively. As observed, speech onsets induce positive "peaks" in $\Delta\mathbf{c}_0$; analogously, speech offsets induce negative peaks. Fig. 5.3 (b) and (c) illustrate the effects of increasing T_{60} on speech offset regions (e.g., between 1.75-2.25 s); the plots correspond to $T_{60} = 0.4$ s and 1 s, respectively. As can be seen, as T_{60} increases $\mathbf{c}_0$ decays at a slower rate, which in turn, decreases the log-energy rate of change. Moreover, due to temporal smearing, the intervals between phonemes are filled with reverberant energy

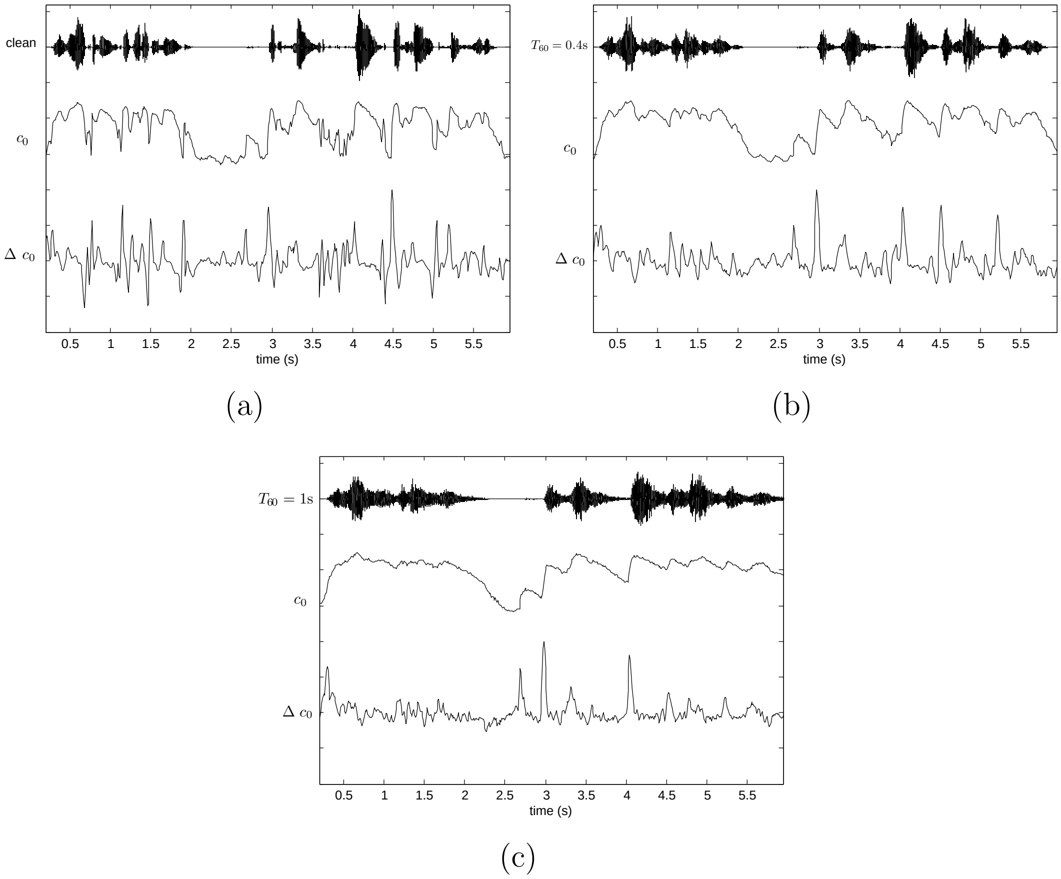


Figure 5.3: From top to bottom: waveform, c_0, and Δc_0, for (a) clean speech, (b) reverberant speech with $T_{60} = 0.4s$, and (c) 1 s.

(e.g., between 0.5-1.75 s), thus also decreasing the log-energy rate of change.

In order to capture such reverberation tail effects, sample statistics are computed from N Δc_0 samples (d_i). In particular, standard deviation (σ_Δ), skewness ($\mathcal{S}_\Delta$), kurtosis ($\mathcal{K}_\Delta$), and median absolute deviation ($\mathcal{D}_\Delta$) are computed according to

$$
\begin{aligned}
\sigma_\Delta &= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i-\bar{d})^2}, &(5.6)\\
\mathcal{S}_\Delta &= \frac{\sqrt{N}\sum_{i=1}^{N}(d_i-\bar{d})^3}{\left(\sum_{i=1}^{N}(d_i-\bar{d})^2\right)^{3/2}}, &(5.7)\\
\mathcal{K}_\Delta &= \frac{N\sum_{i=1}^{N}(d_i-\bar{d})^4}{\left(\sum_{i=1}^{N}(d_i-\bar{d})^2\right)^{2}} - 3, &(5.8)\\
\mathcal{D}_\Delta &= \mathrm{median}_i(|d_i - \mathrm{median}_j(d_j)|), &(5.9)
\end{aligned}
$$

where $\bar{d}$ indicates the sample average of d_i.

The aforementioned inverse relationship between T_{60} and log-energy rate of change can be observed in the σ_Δ versus T_{60} plots depicted in Fig. 5.4 (solid curve). Moreover, since reverberation tail effects are more pronounced in speech offset intervals, it is expected that with an increase in T_{60}, fewer negative peaks will occur in the Δc_0 sequence. A direct consequence of this effect is the increase in positive skewness $\mathcal{S}_\Delta$, as illustrated in Fig. 5.4 (dashed curve). Note that with our speech data, both speech offsets *and* onsets are severely affected by the reverberation tail for very large reverberation times; hence the decrease in $\mathcal{S}_\Delta$ for $T_{60} = 2$ s. Additionally, it is observed that an increase in T_{60} will result in a shift of the variance to large deviations,

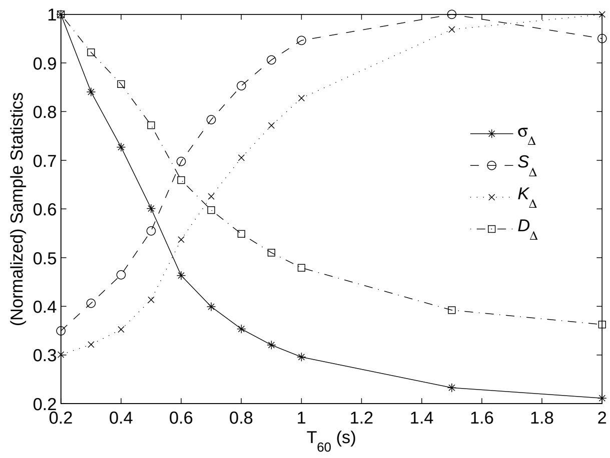


Figure 5.4: Plots of (normalized) sample statistics versus T_{60}. Data points represent average statistics for simulated reverberant speech signals.

rendering the Δc_0 distribution with a heavier tail. Hence, an increase in $\mathcal{K}_\Delta$ is observed, as illustrated in Fig. 5.4 (dotted curve). Lastly, $\mathcal{D}_\Delta$ (dash-dot curve) is used as it provides increased robustness (relative to σ_Δ) to extreme Δc_0 deviations around the mean, an effect commonly observed in multiple-sentence speech signals with an inter-sentence duration that is longer than T_{60}.

Due to the non-linear relationship between T_{60} and Δc_0 sample statistics, we propose to use machine learning algorithms to blindly estimate room acoustical parameters. In our experiments, a support vector regressor (SVR) [96] is used to estimate T_{60}. The input to the SVR is a four-dimensional vector comprised of $\mathbf{u}_s = [\sigma_\Delta, \mathcal{S}_\Delta, \mathcal{K}_\Delta, \mathcal{D}_\Delta]$. As will be shown in Section 5.5.4, a simple adaptation procedure can be used to improve estimation performance in the presence of acoustic background noise.

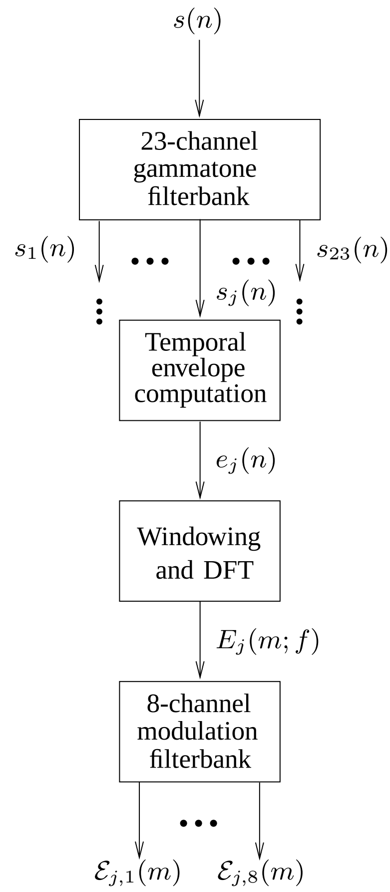


Figure 5.5: Block diagram of the signal processing steps involved in the computation of the spectro-temporal signal representation.

5.4.2 Long-Term Temporal Dynamics

In order to capture long-term temporal dynamics of the reverberant speech signal, we propose to use a spectro-temporal representation of speech termed *modulation spectrum*. The modulation spectrum characterizes the frequency content (or rate of change) of long-term speech temporal envelopes. In our experiments, the spectro-temporal signal representation is obtained using the signal processing steps depicted in Fig. 5.5.

First, the speech signal $s(n)$ is filtered by a bank of critical-band filters. In our

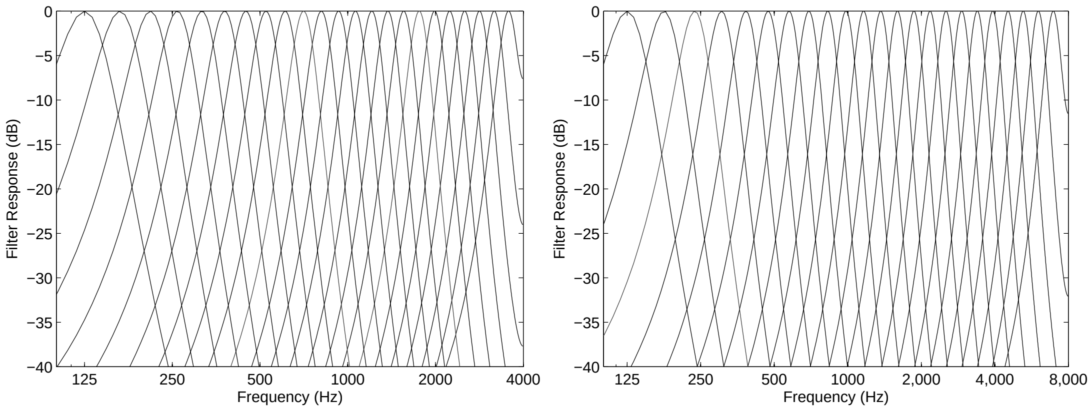


Figure 5.6: Filter responses for the 23-channel gammatone filterbank for (a) narrowband and (b) wideband speech signals.

simulations, a critical-band gammatone filterbank, with 23 filters, is used to emulate the processing performed by the cochlea [153]. Filter center frequencies range from 125 Hz to nearly half the sampling rate (e.g., 3567 Hz for an 8 kHz sampling rate). Filter bandwidths are characterized by the equivalent rectangular bandwidth (ERB) [154]. The ERB for filter j, $j = 1, \ldots, 23$, is given by

$$\mathrm{ERB}_j = \frac{f_j}{Q_{ear}} + B_{min}, \tag{5.10}$$

where f_j represents the center frequency for the filter and Q_{ear} and B_{min} are constants set to 9.265 and 24.7, respectively. The plots in Fig 5.6 (a) and (b) illustrate the frequency response of the 23-channel gammatone filterbank used in our experiments for both narrowband ($F_s = 8$ kHz) and wideband ($F_s = 16$ kHz) speech data, respectively.

The output signal of the j^{th} channel is given by

$$s_j(n) = s(n) * h_j(n), \tag{5.11}$$

where $h_j(n)$ is the impulse response of the j^{th} critical band filter. Temporal dynamics information is obtained from the temporal envelope of $s_j(n)$. In our experiments, the Hilbert transform $\mathcal{H}\{\cdot\}$ is used to obtain temporal envelopes $e_j(n)$. The temporal envelope (also called Hilbert envelope) is computed as the magnitude of the complex analytic signal $\tilde{s}_j(n) = s_j(n) + \jmath\mathcal{H}\{s_j(n)\}$. Hence,

$$e_j(n) = \sqrt{s_j(n)^2 + \mathcal{H}\{s_j(n)\}^2}. \tag{5.12}$$

Temporal envelopes $e_j(n)$ are then multiplied by a 256 ms Hamming window with 32 ms shifts; the windowed envelope for frame m is represented as $e_j(m)$, where the time variable n is dropped for convenience. Here, 256 ms frames are used to obtain long-term temporal dynamics information as well as appropriate resolution for low-frequency modulation frequencies (e.g., around 4 Hz).

The modulation spectrum for critical band j is obtained by taking the discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m)$, i.e., $E_j(m;f) = |\mathcal{F}(e_j(m))|$ where f denotes modulation frequency. Modulation frequency bins are grouped into K bands in order to emulate an auditory-inspired modulation filterbank [155]. The k^{th} modulation band energy for frame m is denoted as $\mathcal{E}_{j,k}(m)$, $k = 1, \ldots, K$. In the experiments described in Section 5.5, $K = 8$ is used as it resulted in superior performance. For the experiments described in Section 5.6, optimal values for K (K^*) are chosen on a per-signal basis. Fig. 5.7 depicts the frequency response of the eight-channel modulation filterbank used in our experiments. Filters are second-order bandpass with quality factor $Q = 2$, as suggested in [155]. Additionally, Table 5.2 reports modulation filter center frequencies (f_c) and filter bandwidths (BW).

The modulation energy $\mathcal{E}_{j,k}(m)$ is then averaged over all active speech frames to

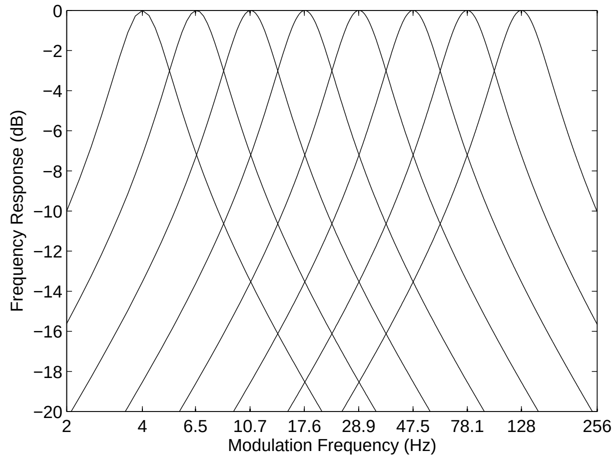


Figure 5.7: Filter responses for the 8-channel modulation filterbank.

Table 5.2: Modulation filter center frequencies (f_c) and bandwidths (BW) expressed in Hz.

	Modulation Frequency Band Index							
	1	2	3	4	5	6	7	8
f_c	4.0	6.5	10.7	17.6	28.9	47.5	78.1	128.0
BW	2.4	3.9	6.5	11.0	18.2	29.1	47.6	78.8

obtain

$$\bar{\mathcal{E}}_{j,k} = \frac{1}{N_{act}} \sum_{i=1}^{N_{act}} \mathcal{E}_{j,k}^{act}(i), \tag{5.13}$$

where N_{act} denotes the number of active speech frames and $\mathcal{E}_{j,k}^{act}(i)$ the modulation energy of such frames; voice activity detection algorithms used in our experiments are described in Section 5.5.5. The $\bar{\mathcal{E}}_{j,k}$ notation will be used throughout the remainder of this chapter to indicate active-speech modulation energy of the j^{th} critical-band signal grouped by the k^{th} modulation filter. A representative illustration of $\bar{\mathcal{E}}_{j,k}$ for a clean speech signal is depicted in Fig. 5.9 (a). Moreover, the notation $\vec{\bar{\mathcal{E}}}_k$ will be used to denote the 23-dimensional energy vector for modulation channel k.

For clean (unreverberated) speech, it is known that Hilbert temporal envelopes contain dominant frequencies ranging from $2-20$ Hz [156, 157] with spectral peaks at approximately 4 Hz, corresponding to the syllabic rate of spoken speech [158]. With reverberant speech, the diffuse IR reverberant tail is often modeled as an exponentially damped Gaussian white noise process [134]. As such, it is expected that reverberant signals attain more Gaussian white-noise like properties with increasing T_{60}. Since the Hilbert envelope can contain frequencies (also termed *modulation frequencies*) up to the bandwidth of its originating signal [159], reverberant signals are expected to contain significant modulation frequency components beyond the $2-20$ Hz range of syllabic modulation frequencies. The plots in Fig. 5.8 assist in illustrating the effects of T_{60} on temporal envelopes. Subplot (a) depicts $e_j(n)$ and the positive portion of $s_j(n)$ ($s_j^+(n)$) for a 256-millisecond frame of clean speech. Subplots (b) and (c), in turn, depict the corresponding signals for reverberant speech with $T_{60} = 0.4$ s and 1 s, respectively. The plots in the figure are for $j = 14$, corresponding to a filter center frequency of 1.2 kHz.

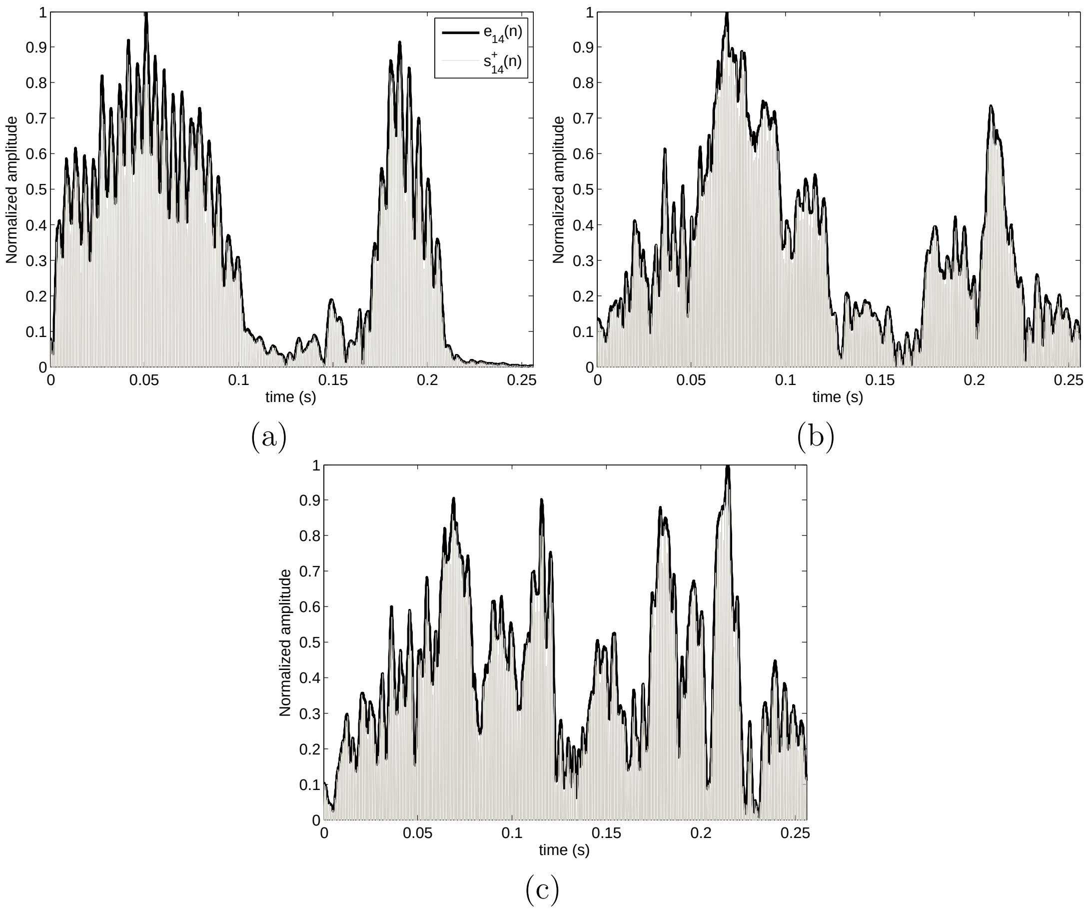


Figure 5.8: Temporal envelope $e_j(n)$ and positive segments of gammatone filtered signal $s_j^+(n)$ for (a) clean speech and reverberant speech with (b) $T_{60} = 0.4s$, and (c) 1 s. The plots are for $j = 14$ corresponding to a filter center frequency of 1.2 kHz.

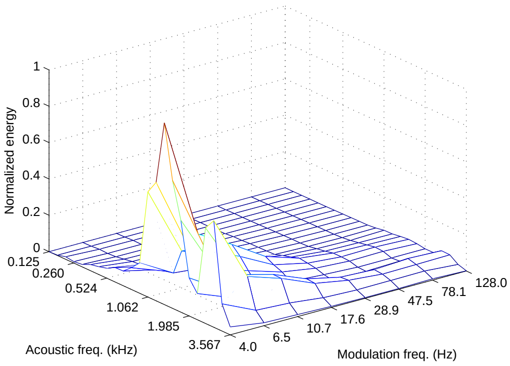
Figure 5.9 depicts the active speech modulation energy $\bar{\mathcal{E}}_{j,k}$ for the speech signals used to produce Fig. 5.8. In the plots, modulation energy values are normalized by the maximum energy obtained over all modulation frequency bands. Fig. 5.9 (a) depicts the normalized modulation energy for a clean speech signal. As observed, most significant modulation frequency components lie below 20 Hz. The plots in Fig. 5.9 (b) and (c), in turn, depict $\bar{\mathcal{E}}_{j,k}$ for the corresponding reverberant speech signals with $T_{60} = 0.4$ s and 1 s, respectively. Increased modulation energy at higher modulation frequency bands is observed for the two plots. Additionally, more pronounced reverberation effects are observed for modulation frequencies greater than 20 Hz (i.e., $k = 5 - 8$).

It can also be observed from Fig. 5.9 that an increase in T_{60} has negligible effect on $\vec{\bar{\mathcal{E}}}_1$, which corresponds to the 4 Hz modulation frequency attributed to the syllabic rate of speech. This insight is used to develop a reverberation-to-speech modulation energy ratio (RSMR) measure computed per modulation frequency channel k and given by

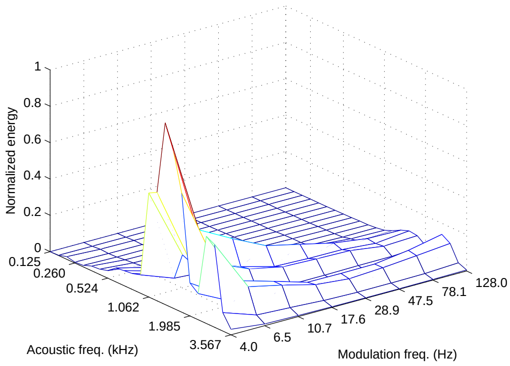
$$\mathrm{RSMR}_k = \frac{\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,k}}{\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,1}}. \tag{5.14}$$

To illustrate the non-linear effects of T_{60} on RSMR, the plots in Fig. 5.10 depict RSMR_k versus T_{60} for $k = 5 - 8$. Data points reflect the average RSMR for the simulated reverberant speech signals described in Section 5.3.3.

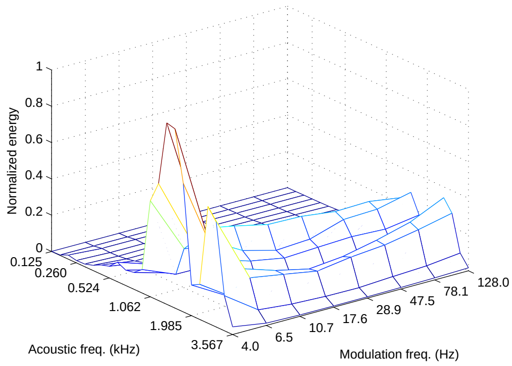
As expected, more pronounced effects are observed for $k = 8$ with an increase in T_{60}. In pilot experiments, we have observed that estimators based only on RSMR_8



(a)



(b)



(c)

Figure 5.9: $\bar{\mathcal{E}}_{j,k}$ for (a) clean speech and reverberant speech with (b) $T_{60} = 0.4$ s, and (c) $T_{60} = 1$ s The gammatone filterbank depicted in Fig. 5.6 (a) is used.

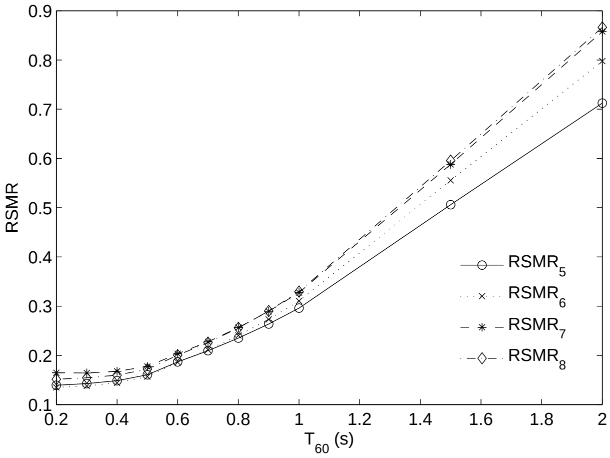


Figure 5.10: Plots of RSMR_k versus T_{60} for $k = 5 - 8$.

attain reliable performance for simulated data but slightly lower performance is attained for reverberant speech generated from recorded room IRs. In order to design estimators that are robust to unseen (real) conditions, a support vector regressor is used to estimate T_{60}. The four-dimensional vector input to the SVR is given by $\mathbf{u}_l = [\mathrm{RSMR}_5, \mathrm{RSMR}_6, \mathrm{RSMR}_7, \mathrm{RSMR}_8]$.

Moreover, as mentioned previously, reverberation tail effects can be quantified from $\vec{\bar{\mathcal{E}}}_k$, $k = 5 - 8$. Speech information, on the other hand, can be obtained from $\vec{\bar{\mathcal{E}}}_1$. This insight is used to compute an overall RSMR measure (ORSMR) which is shown to be highly correlated with DRR. The measure ORSMR is given by

$$\mathrm{ORSMR} = \frac{\sum_{k=5}^{8}\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,k}}{\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,1}} = \sum_{i=5}^{8}\mathrm{RSMR}_i. \tag{5.15}$$

The plot in Fig. 5.11 illustrates a linear regression relationship between ORSMR (expressed in dB) and DRR. Data points represent DRR values described in Table 5.1 and average ORSMR values obtained from English reverberant speech signals generated

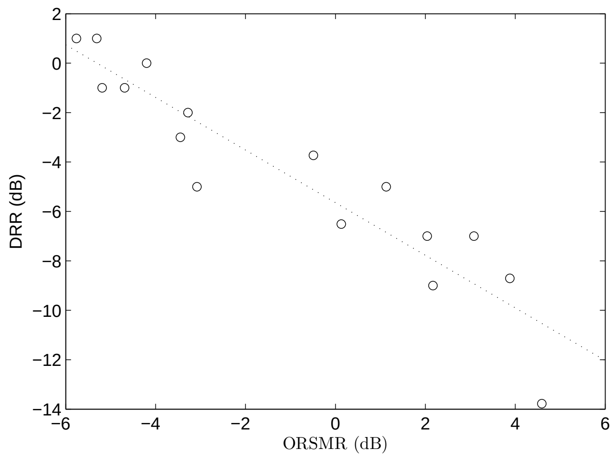


Figure 5.11: Plot of DRR versus ORSMR; the latter is given by (5.15).

with recorded room IRs. Hence, the following DRR estimator ($\widehat{\text{DRR}}$) is proposed:

$$\widehat{\text{DRR}} = -5.6467 - 1.0644 \times \text{ORSMR (dB)}, \tag{5.16}$$

where ORSMR and $\widehat{\text{DRR}}$ are expressed in decibels. In the sequel, the proposed estimators are tested on simulated and recorded reverberant speech.

5.5 Experiments

In this section, experimental setup, performance figures, baseline estimator, and two experiments are described. The first experiment tests the performance of the proposed estimators in reverberant enclosures; the second, in environments corrupted by reverberation *and* acoustic background noise.

5.5.1 Experimental Setup

Reverberant speech signals generated with the SIREAC tool are used to train the support vector regressors. Throughout the remainder of this section, the notation

SVR_s and SVR_l will be used to distinguish blind T_{60} estimators derived from short- and long-term temporal dynamics, respectively. On our data, SVR with radial basis kernels and parameters optimized via linear search are shown to provide the best estimation performance. The results to follow are all based on using radial basis SVR. Additionally, the SIREAC tool is used to generate speech signals degraded by reverberation *and* acoustic background noise. Reverberant speech is generated with T_{60} ranging from 0.2 s to 1 s (with 0.1 s increments) and with babble noise at five signal-to-noise ratio (SNR) levels (5 dB to 25 dB with 5 dB increments). As shown in Section 5.5.4, a simple adaptation process can be used to increase the performance of the proposed T_{60} estimators in the presence of acoustic noise. The "adapted" SVR is termed $\widetilde{\mathrm{SVR}}$ throughout the remainder of this paper.

5.5.2 Performance Figures and Baseline Estimator

Correlation (R), mean square error (MSE), and median absolute error (MAE) are used as estimator figures of merit. The correlation is computed between blindly estimated parameter values (w_i) and parameter measurements obtained from room IR (y_i) using (2.10). The mean square error MSE is given by

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(w_i - y_i)^2, \tag{5.17}$$

and the median absolute error MAE by

$$MAE = \mathrm{median}_i(|w_i - y_i|). \tag{5.18}$$

In the sequel, error measures are reported in milliseconds for T_{60} estimators and in decibels for DRR estimators.

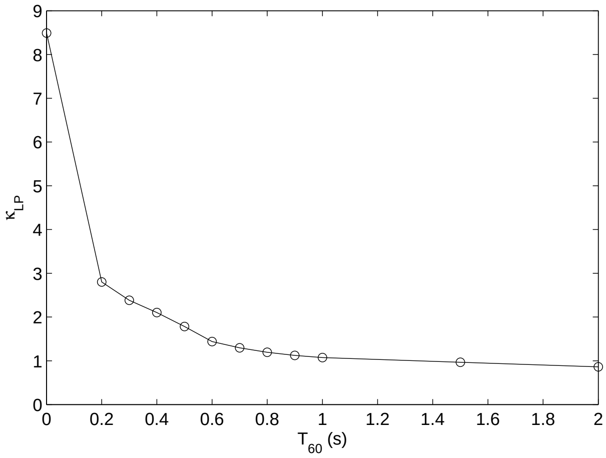


Figure 5.12: Plot of κ_{LP} versus T_{60}. The LP residual kurtosis for clean unreverberated speech is represented as $T_{60} = 0$ in the plot.

The performance of the proposed T_{60} estimators is compared to a baseline estimator based on the kurtosis of 12^{th} order LP residuals (κ_{LP}) computed over 32 ms frames. In our experiments, the LP residual-based method was found to be more robust to background noise when compared to other existing ML-based schemes (e.g., [134, 135]). The plot in Fig. 5.12 shows the non-linear relationship between κ_{LP} and T_{60}. As can be seen, LP residual kurtosis approaches that of a Gaussian distribution with increasing T_{60}. Clean speech, represented by $T_{60} = 0$ s in the plot, attains high κ_{LP}; this is expected as the LP residual for clean speech contains sparse peaks corresponding to the glottal pulses. In our experiments, simulated data is used to train the baseline SVR, henceforth referred to as SVR_{κ}. Moreover, to the best of our knowledge ours is the first blind estimator of DRR, thus comparisons with a baseline are not carried out for $\widehat{\mathrm{DRR}}$.

Table 5.3: Performance comparison of proposed T_{60} estimators for speech corrupted by reverberation.

	Multi-Channel Room IR					Single-Channel Room IR				
	SVR$_\kappa$	SVR$_s$	%	SVR$_l$	%	SVR$_\kappa$	SVR$_s$	%	SVR$_l$	%
R	0.71	0.96	35.2	0.90	26.8	0.82	0.89	8.5	0.86	4.9
MSE (ms)	30.4	11.7	61.5	29.5	3.0	69.9	30.6	56.2	32.2	53.9
MAE (ms)	109.1	80.8	25.9	105.9	2.9	173.5	99.2	42.8	94.7	45.4

5.5.3 Experiment 1 - Reverberation Only

In this experiment, reverberant signals simulated with the SIREAC tool are used to train T_{60} estimators SVR$_s$, SVR$_l$, and SVR$_\kappa$. Bilingual reverberant data generated with real single- and multi-channel recordings of room IR are regarded as unseen data and are used for testing. Table 5.6 reports performance figures for the proposed estimators as well as for the baseline estimator. Columns labeled "%" indicate the percentage increase in R or percentage decrease in MSE and MAE attained with the proposed measures relative to the baseline. As observed, both proposed estimators outperform the baseline method. SVR$_s$ results in superior improvements relative to SVR$_l$ for data generated with the multi-channel room IR. For data generated from the single-channel room IR, both estimators attain similar performance figures, with SVR$_l$ obtaining somewhat lower MAE.

Moreover, as mentioned previously, English reverberant speech data is used to train the coefficients in (5.16). Hence, French reverberant speech data is regarded as unseen and used to test the performance of the proposed DRR estimator. Fig. 5.13 depicts DRR versus average $\widehat{\text{DRR}}$ for the unseen test set; $R = 0.98$, $MSE = 1.11$ (dB), and $MAE = 0.97$ (dB) are attained. The results are encouraging given that no

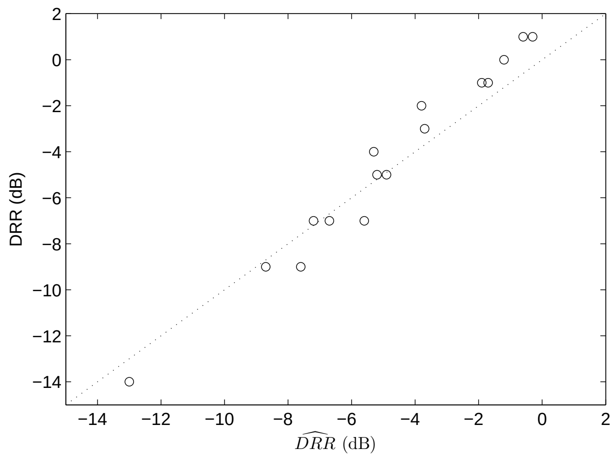


Figure 5.13: Plot of DRR versus average $\widehat{\mathrm{DRR}}$ for unseen French test data.

knowledge of the room IR is used for estimation. Unfortunately, no other signal-based estimators of DRR are available for comparison.

5.5.4 Experiment 2 - Reverberation and Background Noise

To test the performance of the proposed estimators in practical scenarios, we use speech corrupted by reverberation *and* babble (crowd) noise. Table 5.4 reports performance measures for SVR_κ, SVR_s and SVR_l for various noise levels. As can be seen, both proposed estimators outperform the baseline with SVR_l showing reduced sensitivity to noise level. This behaviour is expected as babble noise has speech-like characteristics, thus mostly affecting $\vec{\bar{\mathcal{E}}}_1$. Overall, SVR_s attains average improvements over the baseline of 38.2%, 35.4%, and 23.1% in R, MSE, and MAE, respectively; SVR_l attains average improvements of 22.2%, 81%, and 68.2%.

Despite improved performance over the baseline, high MSE and MAE errors compromise the usability of SVR_s for practical applications. In order to reduce estimation errors, an "adaptation" process is proposed where the estimated SNR is

Table 5.4: Performance comparison of T_{60} estimators for speech corrupted by reverberation and acoustic noise.

	SVR$_\kappa$			SVR$_s$					
SNR (dB)	R	MSE	MAE	R	%	MSE	%	MAE	%
25	0.67	144.1	331.6	0.94	40.3	45.3	68.6	184.4	44.4
20	0.65	192.6	401.1	0.92	41.5	92.2	52.1	273.8	31.7
15	0.63	274.2	498.7	0.88	39.7	185.2	32.5	403.5	19.1
10	0.60	397.2	612.6	0.81	35.0	331.6	16.5	538.3	12.1
5	0.55	551.9	728.2	0.74	34.5	510.1	7.6	669.4	8.1
Average	–	–	–	–	**38.2**	–	**35.4**	–	**23.1**

	SVR$_l$					
SNR (dB)	R	%	MSE	%	MAE	%
25	0.76	13.4	45.9	68.1	146.6	55.8
20	0.76	16.9	46.5	75.9	151.5	62.2
15	0.75	19.0	46.9	82.9	153.7	69.2
10	0.75	25.0	46.7	88.2	153.9	74.9
5	0.75	36.4	55.6	89.9	154.1	78.8
Average	–	**22.2**	–	**81.0**	–	**68.2**

Table 5.5: Performance comparison of adapted T_{60} estimators for speech corrupted by reverberation and acoustic noise.

	$\widetilde{\mathrm{SVR}}_\kappa$		$\widetilde{\mathrm{SVR}}_s$				$\widetilde{\mathrm{SVR}}_l$			
SNR (dB)	MSE	MAE	MSE	%	MAE	%	MSE	%	MAE	%
25	62.8	203.3	32.8	47.8	107.2	47.3	35.0	44.3	117.7	42.1
20	65.6	201.1	39.9	39.2	144.4	28.2	35.2	46.3	114.9	42.9
15	71.1	203.6	46.2	35.0	168.8	17.2	34.8	51.1	120.9	40.6
10	74.8	202.5	52.4	29.9	190.0	6.2	35.2	52.9	119.8	40.8
5	85.7	204.6	56.1	34.5	195.1	4.6	35.9	58.1	126.5	38.2
Average	–	–	–	**37.3**	–	**20.7**	–	**50.5**	–	**40.9**

introduced as an added feature to the support vector estimators. Here, the noise analysis module of the ITU-T P.563 algorithm [50] is used to estimate the SNR. In a controlled experiment, the estimated SNR is shown to be highly correlated with true SNR; $R = 0.96$ is attained. Table 5.5 reports improvements in MSE and MAE for adapted T_{60} estimators; as observed, adaptation substantially reduces estimation errors. Relative to the adapted baseline, $\widetilde{\mathrm{SVR}}_s$ attains average improvements of 37.3% in MSE and 20.7% in MAE. $\widetilde{\mathrm{SVR}}_l$ obtains average improvements of 50.5% and 40.9%, respectively. Improvements in R over the non-adapted estimators are considerably lower – on the order of 7% – for all three estimators, thus are omitted from the table.

5.5.5 Discussion

As can be seen from (5.14) and (5.15), the proposed measures are based on summing per-band modulation energy over the 23 acoustic frequency channels. In order to reduce algorithmic processing time, the critical-band gammatone filterbank can be omitted and per-band modulation energy can be computed over the entire 4 kHz signal bandwidth. On our data, such a simplified configuration is capable of reducing algorithmic processing time by a maximum of 40%. It has been observed, however, that the reduced-complexity configuration lowers measurement performance by as much as 20%, in particular for noise-corrupted environments and for enclosures with low T_{60} (≤ 0.3 s). As a consequence, the reduced-complexity alternative should be considered only if limited resources are available. Moreover, as will be described in Section 5.6, the critical-band gammatone filterbank is useful for objective assessment of perceived reverberation effects, thus has been kept in our experiments.

Additionally, we have experimented with two VAD algorithms. The first is available in the ITU-T G.729 speech codec [127] and the second in the adaptive multi-rate (AMR) wireless speech codec [88]. For reverberant speech files used in Experiment 1 (Section 5.5.3), both VAD algorithms attained similar detection performance. On the other hand, for noise corrupted speech files used in Experiment 2 (Section 5.5.4), the AMR VAD attained improved detection performance, as expected. Notwithstanding, for the purpose of blind room acoustics characterization, similar T_{60} measurement performance is attained with either VAD algorithm, thus signalling the robustness of the proposed measures to voice activity detection errors.

5.6 Quality Measurement for Reverberant and Dereverberated Speech

Recently, several double-ended objective quality measures were tested as estimators of subjective perception of colouration (COL), reverberation tail effects (RTE), and overall quality (OVRL). It was reported that most measures attained poor correlation with subjective listening quality scores ($R \leq 0.40$), and the reverberation decay tail measure attained the highest correlation ($R = 0.62$) with respect to RTE [22]. Such poor performance signals the need for more reliable objective quality measures. Here, long-term temporal dynamics information is investigated for single-ended objective measurement of perceived (de)reverberation effects.

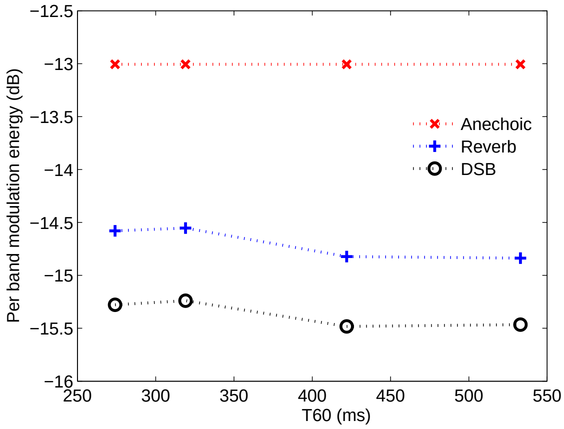
5.6.1 Dereverberation Effects on the Modulation Spectrum

As shown in Section 5.4.2, due to the reverberation tail effect, increased modulation energy is observed for higher frequency modulation channels. To verify the effects of multi-channel *dereverberation* on the modulation spectrum, reverberant speech is generated by convolving 330 anechoic source speech signals with room impulse responses measured by a linear microphone array in four different enclosures (T_{60} values of 274, 319, 422, and 533 ms); a delay-and-sum beamforming dereverberation algorithm is used. The plots in Fig. 5.14 (a)-(b) depict the average per-modulation band energy $\bar{\mathcal{E}}_k$ given by

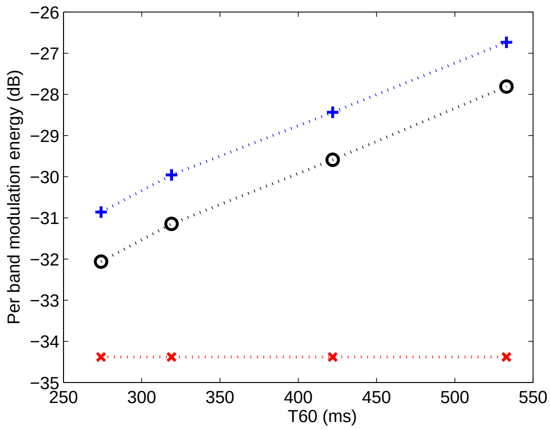
$$\bar{\mathcal{E}}_k = \frac{1}{23}\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,k}, \tag{5.19}$$

averaged over all signals, for modulation bands $k = 1$ and $k = 7$, respectively. The plots depict modulation band energy of anechoic, reverberant, and dereverberated speech processed by the delay-and-sum beamformer (represented by "DSB" in the figure).

As seen from subplot (a), the modulation energy at low modulation frequencies is reduced for reverberant and dereverberated speech signals. Such effects, however, are shown to be relatively *independent* of T_{60} and are likely due to early reflections.



(a)



(b)

Figure 5.14: Per-band modulation energy versus T_{60} for modulation frequency band (a) $k = 1$, and (b) $k = 7$.

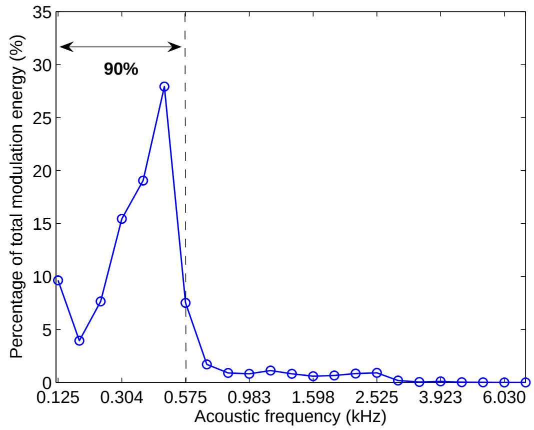
On the other hand, reverberation time dependency is observed for higher frequency modulation channels. From subplot (b), it can be seen that modulation energy increases almost linearly with T_{60}. Moreover, the delay-and-sum beamformer is shown to reduce high-frequency modulation energy by approximately 1 dB relative to reverberant speech. Such gains, however, are quite modest, as an approximate 6.5 dB difference remains between anechoic and dereverberated speech for $T_{60} = 533$ ms.

As mentioned previously, modulation frequency content for acoustic frequency band j is upper-bounded by the bandwidth of critical-band filter j. Hence, speech signals with different acoustic frequency content, subjected to the same quality-degrading reverberation effects, may result in different modulation spectra. Plots in Fig. 5.15 (a)-(b) illustrate one such example; subplots depict the percentage of modulation energy present per acoustic frequency band for speech signals from two different speakers with a reverberation time of 319 ms. As can be seen, for subplot (a) 90% of the total energy is obtained below 575 Hz, whereas for subplot (b) 90% of the total energy is obtained below 983 Hz. The bandwidths of the gammatone filters centered at such frequencies are 86 Hz and 131 Hz, respectively. Hence, according to Fig. 5.7 and Table 5.2, negligible energy at modulation frequency band $k = 8$ is expected from the signal represented in subplot (a).

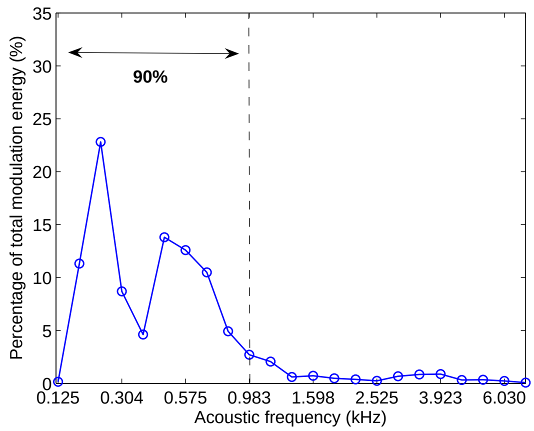
Using this insight, an "adaptive" measure termed speech to reverberation modulation energy ratio (SRMR) is proposed for single-ended quality measurement of reverberant and dereverberated speech. The measure is given by

$$\mathrm{SRMR} = \frac{\sum_{k=1}^{4} \bar{\mathcal{E}}_k}{\sum_{k=5}^{K^*} \bar{\mathcal{E}}_k} \tag{5.20}$$

and is adaptive as the upper summation bound K^* in the denominator is dependent on the speech signal under test. In our simulations, K^* is chosen on a per-signal basis and



(a)



(b)

Figure 5.15: Percentage of modulation energy per acoustic frequency band, for speech signals from two different speakers.

depends on the bandwidth of the lowest gammatone filter for which 90% of the total modulation energy is accounted for. As examples, for the speech signals represented in Fig. 5.15 (a)-(b), $K^* = 7$ and $K^* = 8$, would be used, respectively. To test the performance of the proposed SRMR measure, a subjectively scored reverberant speech database is used.

5.6.2 MARDY Database Description

A subjectively scored multi-channel acoustic reverberation database (MARDY) [22] is used in our experiments. The database uses room impulse responses that were collected with a linear microphone array in an anechoic chamber with reflective panels and with absorptive panels in place. Speaker-to-microphone distances varied from one to four meters (1-meter increments) and T_{60} values ranged from 291 ms to 447 ms. Reverberant speech was generated with the collected room impulse responses and anechoic speech from two speakers (one male and one female); additionally, three dereverberation algorithms were used.

In the experiments described herein, only reverberant speech and speech processed by a conventional delay-and-sum beamformer are used. Speech signals are digitized with 16-bit precision and stored with a 16 kHz sampling rate. More detail regarding the development of the MARDY database can be found in [22]. A subjective listening test was performed following the guidelines described in [19]. In the test, 26 normal hearing listeners rated the subjective perception of colouration (COL), reverberation tail effect (RTE), and overall speech quality (OVRL). Listeners used a 5-point scale where a rating of 5 indicated the best score and a rating of 1 the worst score for a given category. Calibration speech examples were presented to the listeners in order to assist in identification and quantification of colouration and reverberation tail effects.

5.6.3 Experimental Results

The performance of the proposed measure is compared to that of four state-of-the-art algorithms: two are double-ended, ITU-T wideband PESQ (W-PESQ) [160] and PEMO-Q [161], and two are single-ended, ITU-T P.563 [50] and the American National Standards Institute (ANSI) ANIQUE+ algorithm [78]. For the P.563 and ANIQUE+ algorithms, a downsampled (8 kHz) version of the MARDY database is required. We experiment with the narrowband and wideband gammatone filterbanks depicted in Fig. 5.6; the latter is used as it attained somewhat improved performance.

Table 5.6 reports correlation values (R) attained between subjective scores and quality scores obtained with the four quality measurement algorithms and the proposed SRMR measure. Additionally, to demonstrate the gains obtained with the adaptive SRMR measure, a comparison is also carried out with a non-adaptive measure. Denoted by SRMR* in the table, the non-adaptive version uses a fixed $K^* = 8$ value for all speech signals. The column labeled "$\%R \uparrow$" lists the "R-improvement," given by (2.12), obtained by using the proposed SRMR measure.

As observed, the proposed measure is shown to reliably estimate the three quality dimensions for both reverberant and dereverberated speech. Overall, SRMR is shown to outperform state-of-the-art double- and single-ended algorithms by an average 53%, 46%, and 42% for COL, RTE, and OVRL, respectively. Additionally, improvements in performance of 36%, 17%, and 12% are attained relative to SRMR* for all data; more significant gains are obtained for dereverberated speech data. ANIQUE+ is shown to slightly outperform SRMR in OVRL prediction for reverberant speech.

Table 5.6: Performance comparison between SRMR, SRMR*, W-PESQ, PEMO-Q, P.563, and ANIQUE+ on MARDY database. Average improvement is computed over the four quality measurement algorithms.

	Overall (reverberant+dereverberated)						Reverberant					
Algorithm	COL	$\%R\uparrow$	RTE	$\%R\uparrow$	OVRL	$\%R\uparrow$	COL	$\%R\uparrow$	RTE	$\%R\uparrow$	OVRL	$\%R\uparrow$
SRMR	0.82	–	0.83	–	0.80	–	0.81	–	0.84	–	0.81	–
SRMR*	0.73	36.2	0.80	16.6	0.77	12.1	0.73	28.8	0.83	5.9	0.81	0.0
W-PESQ	0.66	48.3	0.81	8.9	0.72	26.0	0.66	44.1	0.82	11.4	0.70	37.3
PEMO-Q	0.61	55.6	0.53	64.1	0.48	61.2	0.70	37.4	0.61	59.9	0.56	56.9
P.563	0.44	68.7	0.46	68.4	0.35	68.6	0.38	69.5	0.41	73.4	0.31	72.7
ANIQUE+	0.72	38.2	0.70	42.6	0.77	12.2	0.77	17.9	0.76	34.7	0.84	-15.3
Average	–	52.7	–	46.0	–	42.0	–	42.2	–	44.9	–	37.9

	Delay-and-sum					
Algorithm	COL	$\%R\uparrow$	RTE	$\%R\uparrow$	OVRL	$\%R\uparrow$
SRMR	0.85	–	0.83	–	0.79	–
SRMR*	0.72	45.8	0.75	33.7	0.72	22.1
W-PESQ	0.67	55.2	0.83	3.8	0.78	4.0
PEMO-Q	0.52	69.1	0.47	68.7	0.38	65.5
P.563	0.54	67.7	0.50	66.4	0.40	64.0
ANIQUE+	0.67	54.7	0.57	61.5	0.67	34.5
Average	–	61.7	–	50.1	–	42.0

Nonetheless, the capability of the proposed measure to reliably estimate colouration and reverberation tail effects, in addition to overall quality, make it a better candidate for single-ended evaluation of reverberant speech and of dereverberation algorithms.

5.7 Conclusions

In this chapter, temporal dynamics information has been proposed to construct blind estimators of room acoustic parameters, namely, reverberation time and direct-to-reverberation energy ratio. Estimators, based on short- and long-term temporal dynamics information, are shown to outperform a baseline system using reverberant speech data with *and* without the presence of acoustic background noise. Moreover, an adaptive speech-to-reverberation modulation energy measure is proposed and shown to reliably estimate perceived room reverberation effects such as colouration and reverberation tail effects in addition to overall quality. The performance of the proposed measure is compared to that of four state-of-the-art quality measurement algorithms and substantial improvement is observed for both reverberant and reverberation-suppressed speech.

Chapter 6

Quality Measurement for Text-to-Speech Systems

6.1 Preamble

This chapter is compiled from material extracted from a manuscript published in the IEEE Signal Processing Letters [87]. Some of the insights described herein have been reported in the Proceedings of the 2008 Blizzard TTS Challenge Workshop [162] and as a contribution to the International Telecommunications Union [163].

6.2 Introduction

Text-to-speech (TTS) synthesis, as the name suggests, attempts to convert arbitrary input text into intelligible and naturally sounding speech. Earlier applications of TTS systems served mostly as an aid to the visually impaired. Today, TTS systems are also being applied in email and short message service readers, automated directory assistance, foreign language education [9], and assistive and augmentative communications [10]. As new applications emerge, the need to deliver high-quality synthesized speech increases. As such, the demand for methods to evaluate the quality of TTS systems has also risen. Evaluation of synthesized speech, however, is not an easy task as various quality dimensions can be assessed (e.g., naturalness, intelligibility). Commonly, multidimensional subjective listening quality tests, such as the one described in ITU-T P.85 [23], are used. For applications such as TTS system tuning, however, several tests may be required throughout the development process; thus subjective assessment is not feasible and objective quality measurement is needed.

To date, there is no universally accepted signal-based objective quality measure for synthesized speech. Most available measures are for corpus-based concatenative TTS systems where a *natural speech* corpus is available (refer to Section 1.2.1.1). In practice, however, natural speech corpora may not be available (e.g., with vocoder-based TTS systems); in such instances, a "reference-free" signal-based measure is required. As described in Chapter 1, reference-free (i.e., single-ended) quality measurement algorithms have been proposed for *natural* speech (e.g., ITU-T P.563 and ANSI ANIQUE+). To the best of our knowledge, a signal-based reference-free quality measure for *synthesized* speech has yet to emerge. Notwithstanding, the aforementioned single-ended ITU-T and ANSI standard algorithms were tested on synthesized speech transmitted over different telephone channels [52]. While the measures were shown to estimate the effects of the transmission channel, poor estimation of source speech quality was attained, signaling the need for an objective quality measure for synthesized speech.

In this chapter, the first steps towards devising a *general-purpose* reference-free measure for TTS system quality diagnosis are described. In particular, hidden Markov models are used to devise text- and speaker-independent artificial reference models of naturally produced speech-feature behavior. Perceptual features, extracted from synthesized speech, are then assessed against gender-dependent reference models by means of a normalized log-likelihood measure. The degree of "consistency" with the reference models is proposed as a measure for multidimensional quality diagnosis. The remainder of this chapter is organized as follows. Section 6.3 describes the signal processing steps needed to compute the proposed quality measure. Section 6.4 describes the databases used in our tests and reports the experimental results. Conclusions are presented in Section 6.5.

6.3 Proposed HMM-Based Quality Measure

The signal processing steps involved in the computation of the proposed HMM-based quality measure are depicted in Fig. 6.1. Pre-processing is first performed to match the characteristics of the signals used to develop the reference models. Voice activity detection (VAD) is then performed on the pre-processed speech signal to remove silence intervals longer than an empirically set value. The feature extraction module serves to compute perceptual and prosodic features; the latter are used to identify talker gender. Pilot experiments have suggested that improved performance is attained if gender-dependent reference models are used. Lastly, perceptual features are assessed against offline-obtained reference hidden Markov models of natural speech-feature behavior via a normalized log-likelihood measure. A detailed description of the signal processing steps is given in the subsections to follow.

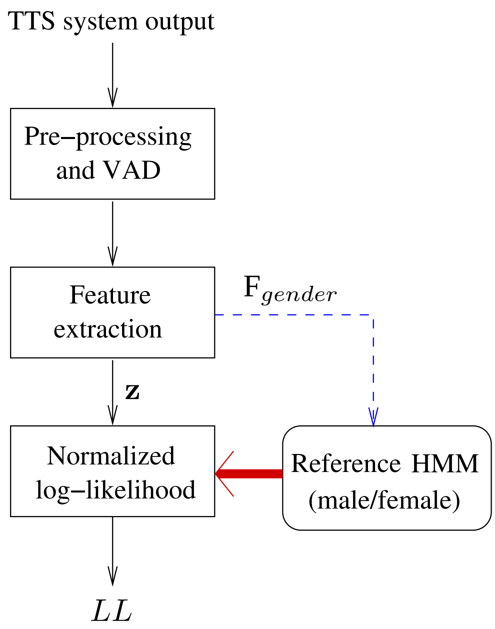


Figure 6.1: Signal processing steps involved in the computation of the proposed HMM-based quality measure. Separate hidden Markov reference models of natural speech-feature behavior are used for male and female speech.

6.3.1 Pre-Processing, VAD and Feature Extraction

In order to match the characteristics of the signals used to train the reference models, pre-processing is applied to the TTS system output. Representative pre-processing steps can include resampling, filtering, and/or signal level normalization. In our experiments, pre-processing consists of bandpass filtering according to [164], down-sampling to 8 kHz, and level normalization to -26 dBov (dB overload) using the P.56 speech-level meter [114]. Moreover, since we are interested in measuring the quality of the output of a TTS system, only active speech segments are analyzed. In our experiments, a simple energy thresholding VAD algorithm is used to remove silence intervals longer than 75 milliseconds; such duration is empirically chosen so as to avoid "artificial" discontinuities introduced by possible VAD errors.

Perceptual features are then computed from active speech; features include 12^{th} order mel-frequency cepstral coefficients (MFCC). The notation $\mathbf{c}_m = \{c_{i,m}\}_{i=0}^{12}$ is used to represent MFCC computed for speech frame m. In our experiments, MFCCs are computed using 25 millisecond windows and 10 millisecond shifts. The zeroth order cepstral coefficient $c_{0,m}$ is used as a log-energy measure. A basic assumption used in this study is that, for natural speech, abrupt changes in signal energy do not occur. Such discontinuities, however, can occur in, e.g., speech produced by a concatenative TTS system. In order to quantify signal energy dynamics, we compute the zeroth delta-cepstral coefficient $\Delta c_{0,m}$, which has been shown useful for temporal discontinuity detection [74]. Feature $\Delta c_{0,m}$ is appended to $\mathbf{c}_m$ to form $\mathbf{z}_m = [\Delta c_{0,m},\ \mathbf{c}_m]$. In Fig. 6.1, $\mathbf{z}$ constitutes features computed for the N_{act} active frames in the synthesized speech signal, i.e., $\mathbf{z} = \{\mathbf{z}_m\}_{m=1}^{N_{act}}$.

Lastly, the fundamental frequency $F0$ is computed with the pitch tracking algorithm described in [89]. $F0$, averaged over all voiced frames, is used to identify talker gender. In pilot experiments, it has been observed that improved quality measurement performance is attained if gender-dependent reference models are used. Motivated by the work described in [50], $F0 = 160$ Hz is used as a threshold to distinguish between male and female voices. A flag indicating talker gender, represented by F_{gender} in Fig. 6.1, is used to indicate which HMM reference model to use.

6.3.2 HMM Reference Models and Log-Likelihood Computation

Speech temporal dynamics provides important information for the measurement of synthesized speech quality and naturalness. As such, we propose to use hidden

Markov reference models trained on naturally-produced speech. The spectral-temporal information captured by the HMM can be used to quantify differences between e.g., natural word endings and abnormal signal interruptions that may occur with synthesized speech. Reference models are obtained using the perceptual features $\mathbf{z}$ described in Section 6.3.1. Features are extracted from the natural speech data described in Section 6.4.1.1 and two reference models are designed, one for male and one for female speech data.

Hidden Markov models have been widely used in speech processing with applications ranging from automatic speech recognition [165] to noise suppression [20]. The reader is referred to [165–167] for a more comprehensive review of HMMs and their applications. Here, HMMs with 8 states are used and the output distribution of each state consists of a Gaussian mixture density with 16 diagonal-covariance Gaussian components. Model parameters, such as state transition probabilities, initial state probabilities, and output distribution parameters, are computed using the expectation-maximization algorithm summarized in [165]. Perceptual features, extracted from the synthesized signal under test, are then assessed against the reference models via the log-likelihood measure. Log-likelihood values are computed using the so-called forward-backward procedure described in [165]; more detail can be found in [168]. Normalization is performed based on the number of active-speech frames N_{act} in the signal under test. Note that the log-likelihood measure, referred to as $\overline{LL}$ in Fig. 6.1, is analogous to the so-called consistency measure described in (2.5).

6.4 Experiments

In this section, a description of the databases used in our tests and the experimental results are reported.

6.4.1 Database Description

A description of the naturally-produced and synthesized speech databases used in our experiments is given in the subsections to follow. Natural speech is used to train gender-dependent HMM reference models and synthesized speech to assess the performance of the proposed quality measure.

6.4.1.1 Natural Speech – Training Data

In order to develop reference models of natural speech-feature behavior, the Kiel Corpus of German read speech is used. Files from the "Siemens" and "Erlangen" sentence subsets, uttered by two male and two female speakers, are used. Visual inspection of spectrograms and pitch contours was used to select speakers with spectral-temporal characteristics different from those in the synthesized speech database. The files are downsampled to 8 kHz, bandpass filtered according to [164], level normalized to -26 dBov, and VAD-processed. Per-gender files are concatenated to produce approximately one hour and 15 minutes of active speech to train the male and female reference HMMs. It is emphasized that the sentences uttered in the training speech dataset differ from the text used to generate the synthesized speech material.

6.4.1.2 Synthesized Speech – Test Data

The synthesized speech database used in our experiments contains speech material from six "off-the-shelf" TTS systems. Three are commercial systems (AT&T, Proser, and Cepstral) and three are from German academic institutions (TU Dresden, TU Berlin, and University of Bonn). Synthesized speech material is produced from the TTS system online demonstration tool. For quality measurement purposes, this exemplifies the scenario where the natural speech corpus is *not* available. A total of 10 speech samples are generated per TTS system, half for male speakers and half for female. The synthesized speech samples have an average duration of 11 seconds and consist of two utterances separated by a silence interval of approximately two seconds. Speech samples were bandpass-filtered according to [164] and level-normalized to -26 dBov prior to listener presentation.

The listening test closely followed the recommendations in ITU-T P.85 [23] and was performed in a silent listening room at the Institute for Phonetics and Digital Speech Processing at Christian-Albrechts-University of Kiel [169]. Seventeen listeners (10 female, 7 male) participated in the test; all were German students and the age ranged from 20-26. Listeners were given a parallel task and asked to rate the synthesized speech signals using eight quality scales. Of the eight scales used, only five are described in ITU-T P.85. Labels of the eight scales used include: *overall impression* (MOS), *listening effort* (LSE), *comprehension problems* (CMP), *articulation* (ART), *naturalness* (NAT), *prosody similarity with natural speech* (PRO), *continuity/fluency* (CFL), and *acceptance* (ACC). Table 6.1 reports the rating scales for dimensions NAT, PRO, and CFL; scales for the five remaining dimensions are described in Section 1.1.2. More details regarding the database can be found in [169].

Table 6.1: Rating scales used in the listening test not described in [23]. Original wordings in German are reported in [169].

Rating	NAT	PRO	CFL
5	Very natural	Very similar	Very fluent
4	Natural	Similar	Fluent
3	Neutral	Somewhat similar	Neutral
2	Unnatural	Dissimilar	Discontinuous
1	Very unnatural	Very dissimilar	Very discontinuous

6.4.2 Experiment Results

To test the performance of the proposed quality measure, Pearson correlation R, attained between $\overline{LL}$ and the various quality dimensions, is used and computed according to (2.10). Table 6.2 reports "per speech sample" correlation coefficients attained for the eight quality dimensions for male and female speech data, considered either separately or jointly; the latter is represented by the column "overall" in the table. For comparison purposes, correlation coefficients attained with the state-of-the-art ITU-T P.563 algorithm are also reported. It is emphasized, however, that synthesized speech does *not* fall within the recommended scope of the standard P.563 algorithm. Unfortunately, no other signal-based reference-free measures are available for comparison.

As observed from the table, the proposed HMM log-likelihood measure correlates well with several quality dimensions, in particular with MOS, NAT, and CFL. Interestingly, $\overline{LL}$ computed for male speech obtains considerably higher correlation values, relative to female speech, for quality dimensions CMP and ART. In turn, higher correlation is attained with female data for dimension PRO. Relative to P.563, substantially higher correlation values are attained with the proposed $\overline{LL}$ measure.

Table 6.2: Performance comparison between $\overline{LL}$ and ITU-T P.563 on eight synthesized speech quality dimensions.

Quality	Proposed $\overline{LL}$			ITU-T P.563		
dimension	Male	Female	Overall	Male	Female	Overall
MOS	0.81	0.72	0.77	0.58	-0.05	0.24
LSE	0.72	0.64	0.65	0.50	0.02	0.20
CMP	0.70	0.45	0.54	0.42	-0.11	0.05
ART	0.74	0.47	0.55	0.53	-0.06	0.11
NAT	0.81	0.80	0.81	0.48	-0.06	0.24
PRO	0.54	0.72	0.61	0.28	-0.18	0.12
CFL	0.74	0.81	0.74	0.51	0.06	0.24
ACC	0.65	0.71	0.67	0.35	-0.10	0.15

Note also that poor correlations are attained with P.563 for female synthesized speech; such intriguing behavior has also been reported in [162, 163] for synthesized speech transmitted over noisy telephone channels.

Furthermore, the work described in [170] suggests cross-gender differences in the subjective perception of synthesized speech quality. In an attempt to compensate for such listener rating "biases," a monotonic polynomial mapping function is applied between $\overline{LL}$ and the subjective quality scores. Monotonic mappings perform scale adjustments but do not alter the ranking of the estimated scores. Table 6.3 reports correlation coefficients attained *after* third-order polynomial regression. As can be seen, a slight improvement in performance is attained after regression; for P.563 predictions, poor correlations remain for female speech.

Ultimately, the aim in objective quality measurement is to develop a measure that ranks similarly with subjective quality ratings. To this end, Spearman rank correlation (R_S) is computed and used as an additional figure of merit. On our data, the proposed $\overline{LL}$ measure attains $R_S = 0.76$ and $R_S = 0.70$ for male and female data,

Table 6.3: Performance comparison between $\overline{LL}$ and ITU-T P.563 after third-order polynomial regression.

Quality	Proposed $\overline{LL}$		ITU-T P.563	
dimension	Male	Female	Male	Female
MOS	0.83	0.74	0.65	0.05
LSE	0.74	0.70	0.59	0.07
CMP	0.72	0.56	0.48	0.02
ART	0.78	0.57	0.62	0.01
NAT	0.84	0.83	0.59	0.20
PRO	0.61	0.72	0.39	0.07
CFL	0.79	0.82	0.61	0.07
ACC	0.70	0.73	0.47	0.03

respectively, for quality dimension MOS. For comparison, P.563 attains $R_S = 0.57$ and $R_S = 0.03$, respectively.

6.4.3 Discussion

While the proposed measure is shown to correlate well with several quality dimensions, it is inferred that further performance gains can be attained if additional features are used in combination with $\overline{LL}$. Representative features can include the mean cepstral deviation ($\bar{\sigma}$) – used as a measure of spectral flatness – which has also been shown useful for spoken dialogue system evaluation [79]. On our data, mean cepstral deviation attains correlation values of -0.64, -0.62, and -0.61 with LSE, CMP, and NAT, respectively (for female speech). Moreover, a sharp decline measure, similar to the one described in [50], is shown to attain correlation values of -0.56, -0.57, and -0.62 with CMP, PRO, and CFL, respectively (for male speech). Feature combination, however, requires access to multiple subjectively scored speech databases in order to optimize feature weights, hence is left for a future study.

6.5 Conclusion

This chapter has described an initial effort at developing a general-purpose single-ended measure for text-to-speech system quality diagnosis. The proposed measure is based on text- and speaker-independent hidden Markov reference models of naturally produced speech and is shown to attain promising results on a multidimensional quality prediction test for both male and female synthesized speech.

Chapter 7

Discussion

In this chapter we discuss and summarize the main results in Chapters 2-6 and their contributions to objective speech quality measurement research and to alternate research fields.

7.1 General-Purpose Speech Quality Measurement

In Chapter 2, a general purpose quality measurement algorithm is constructed from models of speech signals, including clean and degraded speech, and speech corrupted by multiplicative noise and temporal discontinuities. The algorithm has redefined the performance envelope of existing schemes, as it has pioneered the use of:

1. Gaussian mixture densities to model the normative behaviour of speech features, thus allowing for accurate *low complexity* speech quality measurement;

2. Algorithms to detect and quantify spectral flatness related distortions commonly encountered with logarithmically companded PCM, ADPCM, and various other

waveform speech coders;

3. Algorithms to detect and quantify temporal discontinuity distortions commonly encountered with VoIP communications and with speech communications that involve voice activity detection;

4. Advanced pattern recognition algorithms to judiciously combine, using hard or soft decisions, the contributions of the detected distortions to overall speech quality.

Moreover, the algorithm described in Chapter 2 makes use of reference GMMs for clean speech as well as speech degraded by different transmission and/or coding schemes. With modern speech communications, however, signals are subjected to various different sources of degradation, each with its own peculiar impairment to voice quality. As such, if degradation sources can be accurately identified, appropriate degraded speech reference models can be used to improve quality measurement performance. In [81], perceptual features and Gaussian mixture models are investigated for the classification of four modern degradation sources: acoustic background noise, packet loss concealment artifacts, low bitrate coding artifacts, and codec tandeming artifacts. On an unseen test set, the proposed classifier is shown to attain a 98.9% correct identification accuracy [81].

For real-time quality monitoring purposes, knowledge of the degradation source can be used to improve speech quality measurement performance. Due to the modular architecture of the algorithm described in Chapter 2, degradation classification can be easily incorporated to allow for "degradation classification-assisted" speech quality measurement, as depicted in Fig.7.1. The research described in [82] proposes to use degradation-specific reference GMMs and MOS mapping functions for enhanced

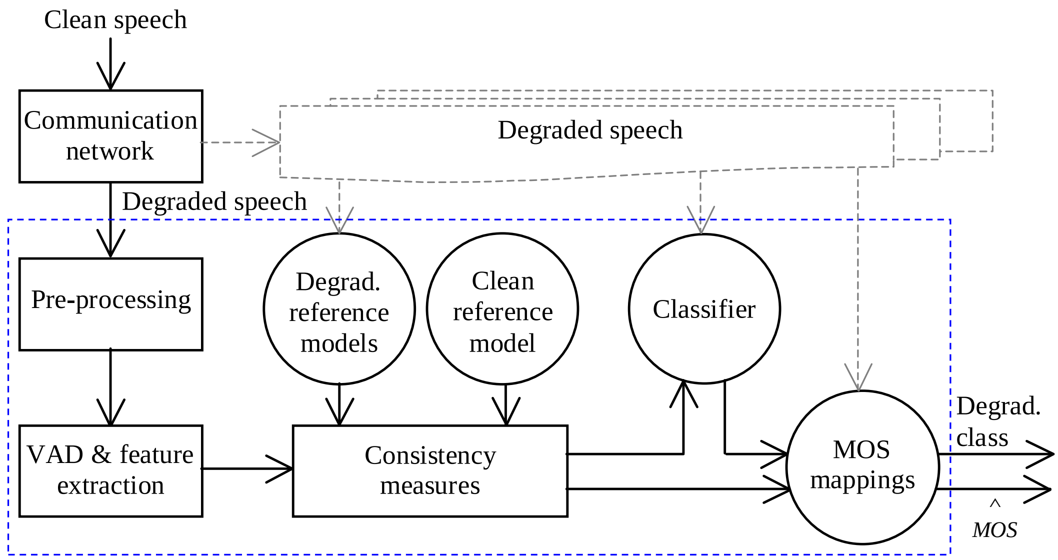


Figure 7.1: Architecture of degradation classification-assisted speech quality measurement algorithm.

speech quality measurement. Experiments show an increase of 10.2% in R and a decrease of 17.6% in $RMSE$ relative to using a "global" degraded-speech reference model, as proposed in Chapter 2. Degradation classification can be used not only to improve quality measurement performance, but can also be used for network diagnosis purposes. In fact, ITU-T has recently initiated an effort to standardize an improved speech quality measurement algorithm termed "objective listening quality assessment," or P.OLQA [171]. The improved algorithm is expected to provide an optional functionality of degradation identification and classification.

7.2 Noise-Suppressed Speech Quality Measurement

In Chapter 3, two architectures are proposed for quality measurement of noise suppressed speech. The first configuration consists of a *network-distributed* speech quality measurement architecture that subsumes existing single- and double-ended quality measurement paradigms. The method improves on current double-ended architectures as it allows for degraded input signals and output signals with quality better than the input, thus, is equipped to handle both quality degradations *and* quality enhancements. Moreover, the proposed architecture allows for diagnosis of the system under test as well as characterization of noise suppression performance. All of the aforementioned functionalities are not available with existing double- or single-ended paradigms.

The second configuration builds on the work described in Chapter 2 to propose a low complexity single-ended quality measurement algorithm for noise suppressed speech. Experiments have shown that in a perceptual speech feature domain distances to reference models of clean, noisy, and noise-suppressed speech are indicative of overall quality. In the work described in Section 3.4, distances between models are computed by means of a fast approximation of the Kullback-Leibler distance. Since Kullback-Leibler distances provide no sense of "direction," three models are used to allow for triangulation. With both aforementioned configurations, the proposed algorithms perform *multidimensional* objective speech quality measurement and three quality dimensions, namely, signal distortion, background intrusiveness, and overall quality, are estimated.

Additionally, the insights obtained with the proposed KLD measures can be used for applications other than speech quality measurement. One such application is blind detection of noise and noise suppression artifacts to test the applicability of conventional double-ended algorithms. Using the perceptual feature properties described in Section 3.4.1.2, a simple test-of-concept experiment is conducted. In the experiment, a three-node classification tree [97] is designed to detect whether a signal is noisy or if it has been processed by a noise suppression algorithm. The classification tree is trained using KLDs computed between the online derived GMM and the three reference GMMs, for both active and inactive speech frames. The designed tree is tested on 96 *unseen* speech signals: 48 are noise-suppressed signals and 48 are signals corrupted by babble and car noise at 0 dB and 5 dB. In this experiment, all 96 signals were correctly classified.

7.3 Hybrid Measurement for Wireless-VoIP Communications

In Chapter 4, the performance of standard single-ended objective quality measurement algorithms is investigated for speech degradations representative of those present in emerging wireless-VoIP communications. It is shown that current signal-based measurement algorithms are sensitive to different VoIP impairments (e.g., packet loss rates) and produce large per-sample quality estimation errors and error variance. Additionally, link parametric methods are shown to be sensitive to distortions that are not captured by connection parameters, such as those present in wireless communication services (e.g., acoustic noise suppression artifacts). In order to overcome such limitations, a hybrid signal-and-link-parametric measurement algorithm, which combines the strengths of pure signal-based and pure link parametric measurement paradigms, is proposed.

With VoIP communications, pure link parametric methods have gained popularity due to their low computational complexity. In Section 4.6, a "codec-integrated" quality measurement paradigm is described that allows for features and voice activity decisions computed by the speech decoder to be shared with the quality measurement algorithm. With such integrated processing, approximately 90% lower computational processing time is attained relative to the standard signal-based ITU-T P.563 algorithm.

7.4 Quality Measurement for Hands-Free Speech Communications

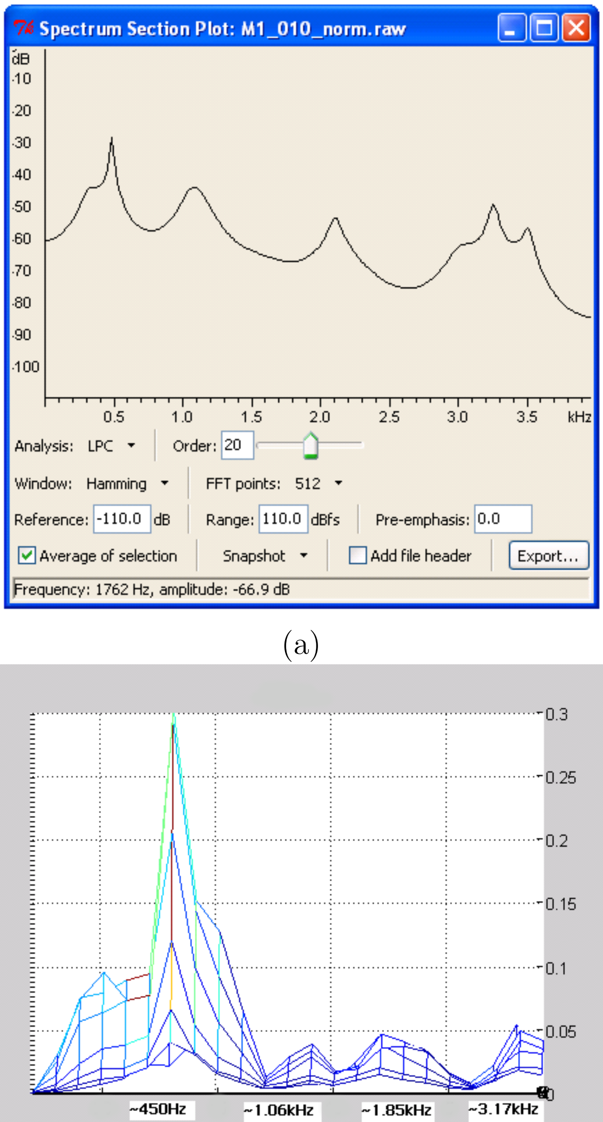
In Chapter 5, several contributions to quality diagnosis for hands-free speech communications are described. With far-field hands-free speech communications, room reverberation acts as a major performance degrading factor for applications such as speech/speaker recognition, pitch tracking, and speaker separation, to name a few. With such applications, if room acoustics characteristics are known beforehand, signal processing strategies can be adapted to improve algorithm performance (see e.g., [172]). In practice, however, *online* estimation of room acoustical parameters is required, thus signal-based measures are needed.

In Section 5.4, short- and long-term temporal dynamics information is used for blind characterization of room acoustics, in particular, to develop estimators of the room reverberation time. Long-term temporal dynamics information is obtained by means of a spectro-temporal signal representation in which speech and reverberation tail effects are shown to be separable. Using this separability property, a signal-based estimator of the room direct-to-reverberation energy ratio is also proposed.

Moreover, as emphasized in [22, 173], current objective speech quality measures attain poor performance when used with reverberant and dereverberated speech signals. In fact, since a reliable objective measure for dereverberated speech is non-existent, researchers currently rely on word error rates produced by state-of-the-art automatic speech recognizers to characterize the "quality" of dereverberated speech [174]. In order to fill this gap, a *multidimensional* objective quality measure is proposed for reverberant and reverberation-suppressed speech. The estimated quality dimensions include colouration, reverberation tail effects, and overall quality. Moreover, the insights obtained from this work have been successfully applied to environment-robust automatic speaker recognition [175], as described below.

Pilot experiments have suggested that energy envelopes obtained across acoustic frequencies, for the first four modulation frequency channels (i.e., in accordance with normative speech behavior, as described in Section 5.4.2), resemble spectral envelopes obtained from higher-order linear prediction (LP) analysis of the speech signal; Fig. 7.2 assists in illustrating this effect. Fig. 7.2 (a) depicts the LP envelope obtained from 20^{th} order LP analysis while Fig. 7.2 (b) depicts the energy envelope obtained across acoustic frequencies for the modulation frequency channel centered at 4 Hz, corresponding to the syllabic rate of spoken speech [158]. As can be seen, visual resemblance is clear and similar peaks and peak positions are found, although not in magnitude. This insight has been used to devise a robust far-field automatic speaker identification (ASI) engine [175].

Today, most state-of-the-art ASI systems are based on mel-frequency cepstral coefficients. As emphasized in [172, 176, 177], however, ASI performance degrades substantially in hands-free far-field applications. Commonly, either dereverberation



(a)

(b)

Figure 7.2: Illustration of similarity between LP envelope and modulation energy envelope. Subplot (a) depicts LP envelope obtained from 20^{th} order LP analysis, and subplot (b) depicts the energy envelope obtained across acoustic frequencies for the first modulation frequency channel.

algorithms or techniques such as cepstral mean subtraction and variance normalization [178] are used prior to speaker recognition in order to improve performance. As described in [175], however, the improvements attained with such techniques are minimal, in particular if the room reverberation time is high ($T_{60} > 0.5$ s). Using the insights described in Section 5.4.2, perceptual features are proposed based on information extracted from the first four modulation frequency channels. Experiments described in [175, 179] serve to demonstrate that an ASI engine based on the proposed spectro-temporal features can outperform an ASI engine based on MFCC by as much as 85% for large rooms with a reverberation time of approximately $T_{60} = 1$ s. An average improvement in identification accuracy of 15% is attained for T_{60} ranging from 0.2 s to 1 s.

7.5 Quality Measurement for Synthesized Speech

In Chapter 6, the first steps towards the development of a reference-free signal-based quality measurement algorithm for synthesized speech are described. Signal-based quality measures available today are for corpus-based concatenative TTS systems where the *natural speech* corpus is available. Such measures, however, are only useful if perceptual degradations are linked to concatenation effects and/or if a reference natural speech corpus is available. Since such requirements are not always met in practice, a reference-free measure is required. In order to fill this gap, the algorithm described in Chapter 2 is modified and adapted for synthesized speech signals. Since temporal dynamics information provides important cues regarding the quality and naturalness of synthesized speech, hidden Markov reference models are used in lieu of Gaussian mixture models. Moreover, in pilot experiments, it is observed that improved performance is attained if gender-dependent reference models are used. Experiments with subjectively scored synthesized speech data, described in Section 6.4, show that the proposed normalized log-likelihood measure attains promising estimation performance for several quality dimensions, in particular dimensions labeled *overall impression, listening effort, naturalness, continuity/fluency*, and *acceptance.*

Chapter 8

Conclusions and Future Research Directions

In this chapter, conclusions of this dissertation are drawn and suggestions for future research directions are presented.

8.1 Conclusions

The evaluation of speech quality is of critical importance in today's technology-mediated speech communications systems, mainly because perceived quality is a key determinant of customer satisfaction. With the fast-paced society we live in, mobility, multi-tasking, and low-cost have become the driving forces behind the advances in wireless, VoIP and hands-free telephony, as well as text-to-speech systems. With these emerging technologies come new sources of degradations and of unwanted perceptual artifacts. With wireless applications, background noise has become a significant impairment and, as a consequence, noise suppression algorithms have gained wide popularity and are present in most recent speech codec standards (e.g., [88, 180]).

With VoIP applications, retransmission is not a viable option, thus packet losses have become a major source of quality degradation. To reduce such distortions, packet loss concealment algorithms are employed. Moreover, with far-field hands-free communications, reverberation and noise have become major quality degradation factors. Due to the adverse effects reverberation has on e.g., automatic speech recognition applications, research into dereverberation algorithms is on the rise. Additionally, while burgeoning text-to-speech synthesis technologies have improved the quality and naturalness of synthesized speech, state-of-the-art systems are still not capable of synthesizing speech that is indistinguishable from naturally produced speech [43].

With the aforementioned technological advancements, users are experiencing new types of distortions and perceptual artifacts. As shown throughout this dissertation, the performance of current state-of-the-art objective speech quality measurement algorithms is compromised for such modern speech communication applications. Since machine-based objective speech quality measurement provides a low-cost means for online quality monitoring and control purposes, more accurate estimators are needed. In this dissertation, several advanced quality measurement algorithms have been proposed and described in detail.

First, a general-purpose speech quality meter is proposed and presented in Chapter 2. The algorithm is based on Gaussian mixture reference models of normative speech behaviour and on innovative techniques to detect and measure multiplicative noise and temporal discontinuities. The algorithm serves as a foundation for the algorithms proposed in subsequent chapters. In Chapter 3, the algorithm is first employed in a network-distributed manner, thus allowing for both quality degradations and enhancements to be handled. It is further expanded to incorporate models of clean, noisy, and noise-suppressed speech, thus allowing for reliable quality measurement for wireless communications with noise suppression capabilities.

In Chapter 4, the algorithm is modified to allow for accurate quality measurement for emerging wireless-VoIP communications. More specifically, a hybrid signal-and-link-parametric measurement paradigm is proposed. Packet switching network parameters are used to estimate a base quality which, in turn, is adjusted according to signal-based distortions measured from the speech signal. The proposed hybrid methodology is shown to overcome the limitations of existing pure signal-based and link parametric algorithms whilst incurring negligible computational overhead.

An alternate scenario, addressed in Chapter 5, is that of far-field hands-free speech communications, where room reverberation acts as a major quality degradation factor. A reverberation-to-speech modulation energy measure is proposed and used for blind characterization of room acoustics. More specifically, the measure is used to derive estimators of the room reverberation time and direct-to-reverberation energy ratio parameters. Furthermore, an adaptive version of the measure is implemented and shown to be a reliable estimator of subjective perception of colouration, reverberation tail effects, and overall quality.

Lastly, a general-purpose quality measurement algorithm for synthesized speech is proposed and described in Chapter 6. Text- and speaker-independent Hidden Markov models, trained on naturally-produced speech, are used to capture normative speech spectral-temporal information. A log-likelihood measure, computed from perceptual features extracted from the synthesized speech signal and the reference models, is proposed and shown to attain promising results on a multidimensional quality test.

8.2 Future Research Directions

- *Wideband Speech Quality Measurement:* Current advances in high-fidelity audio/speech coding and wideband signal extension, combined with the rise of cable and fiber optic networks, will soon allow for widespread use of high-fidelity or *wideband* telephony. Many of the existing algorithms have been optimized for narrowband speech quality measurement. The first attempt to extend ITU-T PESQ to wideband speech quality measurement is discussed in [160]. Tests using wideband PESQ (W-PESQ) are carried out in [181] and it is shown that W-PESQ accuracy is dependent on the speech codec under test. ITU-T efforts are currently under way to standardize a wideband speech quality measurement algorithm [171] and to extend the E-model [67, 182]. As a consequence, wideband speech quality measurement should be an area that will receive significant efforts in the years to come. While the algorithms proposed here have been optimized for narrowband speech, it is believed that the paradigms proposed in Chapters 2-4 and Chapter 6 can be extended to wideband speech. Possible changes include the use of higher-order PLP/MFCC coefficients and retraining of the GMM/HMM reference models; such changes, however, require access to subjectively scored wideband speech data, thus are left for future study. Notwithstanding, as described in Chapter 5, the proposed "wideband" SRMR measure attains accurate quality measurement performance.

- *Quality-Aware Signal Processing and Communications:* Quality-aware signal processing makes use of objective quality measures to systematically adjust algorithm parameters in real-time in order to maximize end-user quality perception. Standard noise suppression algorithm parameters are commonly optimized offline using expert listeners in order to maximize perceptual quality for a given noise type (e.g., street noise). With wireless communications, however, users are mobile and different noise types and levels are experienced throughout the duration of a phone call. It is expected that online adjustment of algorithm parameters will maximize noise suppression and will improve user experience. Dereverberation algorithms, on the other hand, rely on multi-microphone processing or on inverse filtering techniques. Parameters are tuned offline or are adjusted online based on mean-square error optimization. Online parameter adaptation by means of a perceptual quality measure is likely to lead to improved quality. Similarly, quality-aware communications can perform online quality monitoring to control network transmission parameters in order to optimize rate-quality performance. The proposed quality measures can be explored to systematically adjust noise suppression, dereverberation, and text-to-speech algorithm parameters in real-time, as well as to systematically adapt network transmission parameters.

- *Objective Quality Measurement for the Hearing Impaired:* With a rapidly aging population, it is expected that hearing impairments will affect over 20% of the Canadian population by 2020. Currently, user dissatisfaction with commercially available hearing aids is fairly high, thus exacerbating the need for improved signal processing algorithms for the hearing impaired. The design of an objective speech quality measure, tuned to impaired listeners, would be a first step in this direction. The perceptual quality measures proposed in Chapters 2-4 make use of signal processing techniques that emulate the behavior of normal human hearing. In particular, psychoacoustic concepts such as critical band spectral analysis, equal loudness mapping, and intensity-to-loudness power mappings are modeled. For hearing impaired listeners, such precepts are not accurate and need to be updated in order to account for, e.g., sensorineural impairments. Future research into quality measurement for the hearing impaired should focus on adjusting the proposed quality measures to incorporate models of impaired listening. The adapted quality measure can be used to objectively evaluate hearing aids as well as to assist in hearing aid fitting. A longer-term goal can be the development of quality-aware noise suppression algorithms to improve speech reception for hearing aid users in adverse listening environments.

- *Non-Invasive Disordered Speech Quality Diagnosis:* Dysphonia is a disorder of the speech production mechanism in the larynx with perceptual, acoustic, and physical correlates. Persons suffering from dysphonia often experience low self-esteem, shyness, and poor public speaking skills. Speech disorders are commonly diagnosed by means of invasive stroboscopic evaluations and by subjective evaluation of voice production "quality." With the latter, the so-called GRBAS (grade, roughness, breathiness, asthenicity, and strain) test is commonly used where each parameter is scored using a four-point rating scale ranging from 0 denoting normality to 3 denoting extreme pathology. These approaches are time and labour intensive and lack objectivity. Objective measures, in turn, can be used for surgical and/or pharmacological treatment evaluation and for patient rehabilitation monitoring. It is known that the human voice exhibits acoustic evidence of underlying voice disorders through acoustic amplitude fluctuations. As such, amplitude modulation analysis has been used for non-invasive objective speech disorder detection and classification. It is expected that improved performance be attained if spectro-temporal models, such as those used in Chapter 5, are used to exploit amplitude modulations in the voice signal.

- *Objective Image/Video Quality Measurement:* Machine-based algorithms allow computer programs to automate image/video quality measurement in real time, thus playing a crucial role in applications such as compression, steganalysis, and coding. The paradigms proposed here for speech quality measurement are general and can be used for image/video quality measurement. In fact, the first steps have already been taken in [183]. Moreover, the hybrid signal-and-link-parametric quality measurement paradigm can be explored for emerging IP television and video streaming applications.

Bibliography

[1] "Cellphones close the gap," Toronto Star, May 15, 2007.

[2] J.-H. Chen and J. Thyssen, "The broadvoice speech coding algorithm," in *Proc. Intl. Conf. Acoustics, Speech, Signal Processing*, vol. 4, April 2007, pp. 537–540.

[3] Y. Ha, B. Cho, and Y. Yoon, "New regulatory issues for the wireless/mobile VoIP service in Korea," in *Proc. Conf. Technology Management for the Global Future*, vol. 5, July 2006, pp. 2041–2046.

[4] J. Gibson and B. Wei, "Tandem voice communications: Digital cellular, VoIP, and voice over Wi-Fi," in *Proc. Global Telecommunications Conf.*, 2004, pp. 617–621.

[5] P. Perala and M. Varela, "Some experiences with VoIP over converging networks," in *Proc. Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, June 2007.

[6] "Changing the telephone landscapes," CBC News, Nov. 15, 2006. [Online]. Available: http://www.cbc.ca/news/background/voip/

[7] "The cost of calling from abroad," BBC News, Jan. 18, 2007. [Online]. Available: http://news.bbc.co.uk/1/hi/technology/6271621.stm

[8] M. Hatler, D. Phaneuf, and M. Ritter, "Muni wireless broadband: Service oriented mesh-ups," ON World Report, July 2007.

[9] D. Gonzalez, "Text-to-speech applications used in EFL contexts to enhance pronunciation," *Electronic Journal of Teaching English as a Foreign Language*, vol. 11, no. 2, Sept. 2007.

[10] H. T. Bunnell, J. McNicholas, J. B. Polikoff, J. Lilley, and G. Oikonomou, "Personalized synthetic speech for AAC devices: The ModelTalker project," in *Proc. American Speech, Language, and Hearing Association Convention*, Nov. 2002.

[11] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications.* Kluwer Academic Publishers, 2000.

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

[13] ITU-T P.800, "Methods for subjective determination of transmission quality," Intl. Telecom. Union, 1996.

[14] ITU-T P.830, "Subjective performance assessment of telephone-band and wide-band digital codecs," Intl. Telecom. Union, 1996.

[15] ITU-T P.800.1, "Mean opinion score (MOS) terminology," Intl. Telecom. Union, 2003.

[16] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. Intl. Conf. on Acoustics, Speech, Signal Processing*, 1977, pp. 204–207.

[17] D. Sen, "Determining dimensions of speech quality from PCA and MDS analysis of the Diagnostic Acceptability Measure," in *Proc. Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, 2001, (3 pages).

[18] M. Waltermann, K. Scholtz, A. Raake, U. Heute, and S. Moller, "Underlying quality dimensions of modern telephone connections," in *Proc. Intl. Conf. Spoken Language Processing*, 2006, pp. 2170–2173.

[19] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," Intl. Telecom. Union, 2003.

[20] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 1846–1856, 1989.

[21] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, July-Aug. 2007.

[22] J. Wen, N. Gaubitch, E. Habets, T. Myatt, and P. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Intl. Workshop Acoustic Echo and Noise Control*, Sept. 2006, (4 pages).

[23] ITU-T P.85, "A method for subjective performance assessment of the quality of speech voice output devices," Intl. Telecom. Union, 1994.

[24] P. Kroon, "Evaluation of speech coders," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, 1995, ch. 13, pp. 467–494.

[25] A. Rix, J. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality – technology and applications," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 6, pp. 1890–1901, Nov. 2006.

[26] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, New Jersey: Prentice-Hall, 1988.

[27] N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, New Jersey: Prentice-Hall, 1984.

[28] R. Kubichek, D. Atkinson, and A. Webster, "Advances in objective voice quality assessment," in *Proc. IEEE Globecom*, 1991, pp. 1765–1770.

[29] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, June 1992.

[30] ITU-T Rec. P.861, "Objective quality measurement of telephone-band (300-3400 hz) speech codecs," Intl. Telecom. Union, Aug. 1996.

[31] S. Voran, "Objective estimation of perceived speech quality – Part I: Development of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 4, pp. 371–382, July 1999.

[32] ——, "Objective estimation of perceived speech quality – Part II: Evaluation of the measuring normalizing block technique," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 4, pp. 383–390, July 1999.

[33] W. Zha and W.-Y. Chan, "Objective speech quality measurement using statistical data mining," *EURASIP Journal of Applied Signal Processing*, vol. 2005, no. 9, pp. 1410–1424, June 2005.

[34] T. H. Falk, W.-Y. Chan, and P. Kabal, "Speech quality estimation using Gaussian mixture models," in *Proc. Intl. Conf. Spoken Lang. Proc.*, Oct. 2004, pp. 2013–2016.

[35] T. H. Falk and W.-Y. Chan, "A sequential feature selection algorithm for GMM-based speech quality estimation," in *Proc. European Signal Proc. Conf.*, Sept. 2005, (4 pages).

[36] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Intl. Telecom. Union, 2001.

[37] S. Pennock, "Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm," in *Proc. Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, Jan. 2002, (20 pages).

[38] M. Varela, I. Marsh, and B. Gronvall, "A systematic study of PESQ's behavior (from a networking perspective)," in *Proc. Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, May 2006, (11 pages).

[39] S. Broom, "VoIP quality assessment: taking account of the edge device," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 6, pp. 1977–1983, Nov. 2006.

[40] ITU-T P.862.3, "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," Intl. Telecom. Union, 2005.

[41] J. Vepa, S. King, and P. Taylor, "Objective distance measures for spectral discontinuities in concatenative speech synthesis," in *Proc. Intl. Conf. Spoken Language Proc.*, Sept. 2002, (4 pages).

[42] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. Intl. Conf. Acoustics, Speech, and Signal Processing*, June 2001, pp. 837–840.

[43] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. Blizzard Challenge Workshop*, 2007, (12 pages).

[44] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. European Congress on Acoustics*, 2005, pp. 2725–2728.

[45] J. Liang and R. Kubichek, "Output-based objective speech quality," in *Proc. IEEE Vehicular Technology Conf.*, vol. 3, June 1994, pp. 1719–1723.

[46] T. H. Falk and W.-Y. Chan, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Proc. Letters*, vol. 13, no. 2, pp. 108–111, Feb. 2006.

[47] T. H. Falk, Q. Xu, and W.-Y. Chan, "Non-intrusive GMM-based speech quality measurement," in *Proc. Intl. Conf. Acoustics, Speech, Signal Processing*, vol. 1, March 2005, pp. 125–128.

[48] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," in *IEE Proc. Vision, Image and Signal Processing*, vol. 147, no. 6, Dec. 2000, pp. 493–501.

[49] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, Sept. 2005.

[50] ITU-T P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," Intl. Telecom. Union, 2004.

[51] L. Malfait, J. Berger, and M. Kastner, "P.563 - The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[52] S. Möller, D.-S. Kim, and L. Malfait, "Estimating the quality of synthesized and natural speech transmitted through telephone networks using single-ended prediction models," *Acta Acustica United with Acustica*, vol. 94, pp. 21–31, 2008.

[53] T. H. Falk, H. Yuan, and W.-Y. Chan, "A hybrid signal-and-link-parametric approach to single-ended quality measurement of packetized speech," in *Proc. Intl. Conf. Acoustics, Speech, Signal Processing*, vol. 4, April 2007, pp. 841–844.

[54] L. Ding, A. Radwan, M. El-Hennaway, and R. Goubran, "Performance study of objective voice quality measures in VoIP," in *Proc. Symposium on Computers and Communications*, July 2007, pp. 197–202.

[55] T. H. Falk, H. Yuan, and W.-Y. Chan, "Single-ended quality measurement of noise suppressed speech based on Kullback-Leibler distances," *Journal of Multimedia*, vol. 2, no. 5, pp. 19–26, Sept. 2007.

[56] A. Ekman and B. Kleijn, "Improving quality prediction accuracy of P.563 for noise suppression," in *Proc. Intl. Workshop Acoustic Echo and Noise Control*, Sept. 2008.

[57] T. H. Falk, H. Yuan, and W.-Y. Chan, "Spectro-temporal processing for blind estimation of reverberation time and single-ended quality measurement of reverberant speech," in *Proc. Interspeech/Eurospeech Conf.*, Aug. 2007, pp. 514–517.

[58] IETC RFC 3550, "RTP: A transport protocol for real-time applications," July 2003.

[59] IETC RFC 3551, "RTP profile for for audio and video conferences with minimal control," July 2003.

[60] IETC RFC 3611, "RTP Control Protocol Extended Reports (RTCP XR)," Nov. 2003.

[61] N. Johannesson, "The ETSI computation model: a tool for transmission planning of telephone networks," *IEEE Communications Magazine*, vol. 35, no. 1, pp. 70–79, Jan. 1997.

[62] ITU-T Rec. G.107, "The E-model, a computational model for use in transmission planning," Intl. Telecom. Union, 2005.

[63] ITU-T Rec. G.113, "Transmission impairments due to speech processing," Intl. Telecom. Union, 2001.

[64] ITU-T Rec. G.113 - Appendix I, "Provisional planning values for the equipment impairment factor Ie and packet loss robustness factor Bpl," Intl. Telecom. Union, 2002.

[65] ITU-T P.833, "Methodology for derivation of equipment impairment factors from subjective listening-only tests," Intl. Telecom. Union, 2001.

[66] ITU-T P.834, "Methodology for the derivation of equipment impairment factors from instrumental models," Intl. Telecom. Union, 2002.

[67] S. Moller, A. Raake, N. Kitawaki, A. Takahashi, and M. Waltermann, "Impairment factor framework for wideband speech codecs," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 14, no. 6, pp. 1969–1976, Nov. 2006.

[68] Telchemy, "Voice quality estimation in wireless and TDM environments," Application Note. Series: Understanding VoIP Performance, April 2006.

[69] M. Chu and H. Peng, "An objective measure for estimating MOS of synthesized speech," in *Proc. European Conf. Speech Communications and Technology*, 2001, pp. 2087–2090.

[70] A. Clark, "VoIP performance management," in *Proc. Internet Telephony Conf.*, 2005, (14 pages). [Online]. Available: http://www.telchemy.com/Conferences/2005/InteropNYDec2005-Clark.pdf

[71] C. Hoene, S. Wiethlter, and A. Wolisz, "Predicting the perceptual service quality using a trace of VoIP packets," in *Proc. Intl. Workshop on Quality of Future Internet Services*, 2004, pp. 21–30.

[72] L. Sun and E. Ifeachor, "New methods for voice quality evaluation for IP networks," in *Proc. Intl. Teletraffic Congress*, Sept. 2003, pp. 1201–1210.

[73] L. Ding, Z. Lin, A. Radwan, M. El-Hennaway, and R. Goubran, "Non-intrusive single-ended speech quality assessment in VoIP," *Speech Communications*, vol. 49, no. 6, pp. 477–489, June 2007.

[74] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[75] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[76] V. Grancharov, J. Plasberg, J. Samuelsson, and B. Kleijn, "Generalized postfilter for speech quality enhancement," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 16, no. 1, pp. 57–64, Jan. 2008.

[77] D.-S. Kim and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, May 2007.

[78] ATIS-PP-0100005.2006, "Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality," American National Standards Institute, 2006.

[79] S. Möller, K.-P. Engelbrecht, M. Pucher, P. Frölich, L. Huo, U. Heute, and F. Oberle, "TIDE: A testbed for interactive spoken dialogue system evaluation," in *Proc. Intl. Conf. Speech and Computers*, Oct. 2007, (6 pages).

[80] T. H. Falk and W.-Y. Chan, "Enhanced non-intrusive speech quality measurement using degradation models," in *Proc. Intl. Conf. Acoustics, Speech, Signal Processing*, vol. I, May 2006, pp. 837–840.

[81] H. Yuan, T. H. Falk, and W.-Y. Chan, "Classification of speech degradations at network endpoints using psychoacoustic features," in *Proc. Canadian Conf. Electrical and Computer Eng.*, 2007, pp. 1602–1605.

[82] ——, "Degradation-classification assisted single-ended quality measurement of speech," in *Proc. Intl. Conf. Spoken Lang. Proc.*, Aug. 2007, pp. 1689–1692.

[83] T. H. Falk and W.-Y. Chan, "Performance study of objective speech quality measurement for modern wireless-VoIP communications," *Speech Communication*, 2008, in press, (9 pages).

[84] ——, "Hybrid signal-and-link-parametric speech quality measurement for VoIP communications," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 16, no. 8, pp. 1579–1589, Nov. 2008.

[85] ——, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. on Instrumentation and Measurement*, 2008, in press, (12 pages).

[86] ——, "A non-intrusive quality measure of dereverberated speech," in *Proc. Intl. Workshop Acoustic Echo and Noise Control*, Sept. 2008.

[87] T. H. Falk and S. Moller, "Towards signal based instrumental quality diagnosis for text-to-speech systems," *IEEE Signal Proc. Letters*, vol. 15, pp. 781–784, 2008.

[88] 3GPP2 TS 26.094, "Adaptive multi-rate (AMR) speech codec: voice activity detector (VAD), release 6," Dec. 2004.

[89] D. Talkin, *A Robust Algorithm for Pitch Tracking (RAPT).* Elsevier Science Publishers, pp. 495–518, in Speech Coding and Synthesis, 1995. Editors: W. B. Kleijn and K. K. Paliwal.

[90] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoustical Society of America*, vol. 87, pp. 1738–1752, April 1990.

[91] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing.* Prentice-Hall, 2001.

[92] H. Hermansky, "Mel cepstrum, deltas, double-deltas-What else is new?" in *Conf. on Robust Methods for Speech Recognition in Adverse Conditions*, 1999.

[93] N. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," in *Proceedings of the IEEE*, vol. 62, no. 5, May 1974, pp. 611–632.

[94] ITU-T P.810, "Modulated noise reference unit - MNRU," Intl. Telecom. Union, 1996.

[95] S. Voran, "Observations on the t-reference condition for speech coder evaluation," Contribution to CCITT SG-12 Experts Group on Speech Quality, document number SQ.13.92, Feb. 1992.

[96] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 1995, New York.

[97] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees.* Monterey, CA: W & B Publishers, 1984.

[98] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[99] A. Dempster, N. Lair, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[100] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, March 1991.

[101] S. Voran, "Perception of temporal discontinuity impairments in coded speech - Proposal for objective estimators and some subjective test results," in *Proc. Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, May 2003.

[102] J. Canny, "Finding edges and lines in images," MIT – Artificial Intelligence Laboratory, Tech. Rep. 720, 1983.

[103] ——, "A computational approach to edge detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, Nov. 1986.

[104] M. Hansen and B. Kollmeier, "Continuous assessment of time-varying speech quality," *J. Acoustical Society of America*, vol. 106, no. 5, pp. 2888–2899, Nov. 1999.

[105] S. Voran, "A basic experiment on time-varying speech quality," in *Proc. Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, June 2005, (14 pages).

[106] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," Intl. Telecom. Union, Feb. 1998.

[107] L. Thorpe and W. Yang, "Performance of current perceptual objective speech quality measures," in *Proc. IEEE Speech Coding Workshop*, 1999, pp. 144–146.

[108] ITU-T P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO," Intl. Telecom. Union, 2003.

[109] V. Mattila, "Objective measures for the characterization of the basic functioning of noise suppression algorithms," in *Proc. Intl. Conf. on Measurement of Speech and Audio Quality in Networks*, May 2003, (12 pages).

[110] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Intl. Conf. on Spoken Lang. Proc.*, 2006, pp. 1447–1450.

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

[112] E. Paajanen, B. Ayad, and V. Mattila, "New objective measures for characterisation of noise suppression algorithms," in *Proc. IEEE Speech Coding Workshop*, Sept. 2000, pp. 23–25.

[113] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. Intl. Conf. on Acoustics, Speech, Signal Processing*, vol. I, May 2006, pp. 153–156.

[114] ITU-T P.56, "Objective measurement of active speech level," Intl. Telecom. Union, 1993.

[115] T.-H. Hwang, L.-M. Lee, and H.-C. Wang, "Cepstral behaviour due to additive noise and a compensation scheme for noisy speech recognition," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 145, no. 5, pp. 316–321, Oct. 1998.

[116] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Noise-dependent postfiltering," in *Proc. Intl. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, Jan 2004, pp. 457–460.

[117] M. Do, "Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models," *IEEE Signal Proc. Letters*, vol. 10, no. 4, pp. 115–118, April 2003.

[118] Z. Liu and Q. Huang, "A new distance measure for probability distribution function of mixture type," in *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, June 2000, pp. 616–619.

[119] D. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler distance," Rice University, Tech. Rep., 2001.

[120] ITU-T Rec. P.805, "Subjective evaluation of conversational quality," Intl. Telecom. Union, April 2007.

[121] ITU-T Rec. G.711-Annex I, "A high quality low-complexity algorithm for packet loss concealment with G.711," Intl. Telecom. Union, 1996.

[122] ITU-T Rec. G.191, "Software tools for speech and audio coding standardization," Intl. Telecom. Union, 2005.

[123] ITU-T P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," Intl. Telecom. Union, 1996.

[124] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, Feb. 1978.

[125] ITU-T Rec. G.108, "Application of the E-model: a planning guide," Intl. Telecom. Union, 1999.

[126] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "PESQ - the new ITU standard for end-to-end speech quality assessment," in 109^{th} *AES Convention*, Sept., pp. 1–18, pre-print 5260.

[127] ITU-T Rec. G.729 - Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," Intl. Telecom. Union, Nov. 1996.

[128] Z. Zivkovic and F. van der Heijden, "Recursive unsupervised learning of finite mixture model," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 651–656, May 2004.

[129] N. Vlassis and A. Likas, "A greedy EM algorithm for Gaussian mixture learning," *Neural Processing Letters*, vol. 15, pp. 77–87, 2002.

[130] J. Gruber and L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems," *IEEE Trans. on Communications*, vol. 33, no. 8, pp. 801–808, Aug. 1985.

[131] P. Kabal, "ITU-T G.723.1 speech coder: a Matlab implementation," McGill University, Tech. Rep., Aug. 2004.

[132] T. Halmrast, "Sound coloration from (very) early reflections," in *Proc. Meeting Acoustical Society of America*, June 2001, (7 pages).

[133] P. Rubak, "Coloration in room impulse responses," in *Proc. Joint Baltic-Nordic Acoustics Meeting*, June 2004, pp. 1–14.

[134] R. Ratnam, D. Jones, B. Wheeler, W. O'Brien, C. Lansing, and A. Feng, "Blind estimation of reverberation time," *J. Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[135] H. Lollmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proc. Intl. Workshop Acoustic Echo and Noise Control*, Sept. 2008, (4 pages).

[136] M. Wu and D. Wang, "A pitch-based method for the estimation of short reverberation time," *Acta Acustica United with Acustica*, vol. 92, pp. 337–339, 2006.

[137] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Intl. Conf. on Acoustics, Speech, Signal Processing*, May 2001, pp. 3701–3704.

[138] N. Gaubitch, D. Ward, and P. Naylor, "Statistical analysis of autoregressive modeling of reverberant speech," *J. Acoustical Society of America*, vol. 120, no. 6, pp. 4031–4039, Dec. 2006.

[139] B. Yegnanarayana and P. Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[140] E. Habets, N. Gaubitch, and P. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *Proc. Intl. Conf. on Acoustics, Speech, Signal Processing*, March 2008, pp. 4577–4580.

[141] H. Kuttruff, *Room Acoustics*, 4th ed. Elsevier, 2000.

[142] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Society of America*, vol. 65, no. 4, pp. 943–951, April 1979.

[143] W. Sabine, *Collected Papers on Acoustics.* Harvard Univ. Press, 1922.

[144] M. Schroeder, "New method of measuring reverberation time," *J. Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, March 1965.

[145] ISO3382 - Acoustics, "Measurement of the reverberation time of rooms with reference to other acoustical parameters," Intl. Organization for Standardization, 2000.

[146] T. Curtis, "Characterization of room coloration by moments of room spectral response," *J. Acoustical Society of America*, vol. 58, Fall 1975, Suppl. No. 1.

[147] J. Jetztz, "Critical distance measurement of rooms from the sound energy spectral envelope," *J. Acoustical Society of America*, vol. 65, no. 5, pp. 1204–1211, May 1979.

[148] S. Bech, "Timbral aspects of reproduced sound in small rooms," *J. Acoustical Society of America*, vol. 99, no. 4, pp. 3539–3549, 1996.

[149] H. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. of Interspeech Conf.*, 2005, (4 pages).

[150] W. Ward, G. Elko, R. Kubli, and C. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. Wallace Clement Sabine Centennial Symposium*, 1994, pp. 343–346.

[151] L. Couvreur and C. Couvreur, "Blind model selection for automatic speech recognition in reverberant environments," *J. of VLSI Signal Processing*, vol. 36, pp. 189–203, 2004.

[152] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[153] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Apple Computer, Perception Group, Tech. Rep., 1993.

[154] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–38, 1990.

[155] T. Dau, D. Puschel, and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system. I – model structure,” *J. Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[156] R. Drullman, J. Festen, and R. Plomp, “Effect of temporal envelope smearing on speech reception,” *J. Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[157] ——, “Effect of reducing slow temporal modulations on speech reception,” *J. Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[158] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, “Intelligibility of speech with filtered time trajectories of spectral envelopes,” in *Proc. Intl. Conf. Speech and Lang. Proc.*, Oct. 1996, pp. 2490–2493.

[159] Z. Smith, B. Delgutte, and A. Oxenham, “Chimaeric sounds reveal dichotomies in auditory perception,” *Letters to Nature*, vol. 416, pp. 87–90, March 2002.

[160] ITU-T P.862.2, “Wideband extension to Rec. P.862 for the assessment of wideband telephone networks and speech codecs,” Intl. Telecom. Union, 2007.

[161] R. Huber and B. Kollmeier, “PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception,” *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[162] T.H.Falk, S. Moller, V. Karaiskos, and S. King, “Improving instrumental quality prediction performance for the Blizzard Challenge,” in *Proc. Blizzard Challenge Text-to-Speech Workshop*, Sept. 2008, (5 pages).

[163] ITU-T Contribution COM 12-180, "Single-ended quality estimation of synthesized speech: Analysis of the Rec. P.563 internal signal processing," Intl. Telecom. Union, 2008, (Authors: S. Möller and T. H. Falk).

[164] ITU-T Rec. G.712, "Transmission performance characteristics of pulse code modulation channels," Intl. Telecom. Union, 2001.

[165] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[166] M. Gales and S. Young, *The Application of Hidden Markov Models in Speech Recognition*, ser. Foundations and Trends in Signal Processing. Publishers Inc., 2007.

[167] X. Huang, Y. Ariki, and M. Jack, *Hidden Markov Models for Speech Recognition*. Edinburgh Univ Press, 1991.

[168] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Intl. Computer Science Institute, Tech. Rep. ICSI-TR-97-021, 1997. [Online]. Available: http://ssli.ee.washington.edu/people/bilmes/mypapers/em.pdf

[169] K. Seget, "Untersuchungen zur auditiven qualität von sprachsyntheseverfahren (Study of perceptual quality of text-to-speech systems)," July 2007, Bachelor thesis, Christian-Albrechts-University of Kiel.

[170] J. Mullennix, S. Stern, S. Wilson, and C. Dyson, "Social perception of male and female computer synthesized speech," *Computers in Human Behavior*, vol. 19, pp. 407–424, 2003.

[171] ITU-T Study Group 12 Temporary Document TD-42, "Requirements for a new model for objective speech quality assessment P.OLQA," Intl. Telecom. Union, June 2006.

[172] J. Gammal and R. Goubran, "Combating reverberation in speaker identification," in *Proc. Instrumentation and Measurement Technology Conference*, May 2005, pp. 687–690.

[173] E. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, Netherlands, June 2007.

[174] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: experimental validation," *EURASIP Journal on Audio, Speech, and Music Proc.*, 2007, (19 pages).

[175] T. H. Falk and W.-Y. Chan, "Spectro-temporal features for robust far-field speaker identification," in *Proc. Intl. Conf. Spoken Lang. Proc.*, Sept. 2008, pp. 634–637.

[176] P. Castellano, S. Sridharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *Proc. Intl. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 1996, pp. 117–120.

[177] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 7, pp. 2023–2032, Sept. 2007.

[178] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 15, no. 7, pp. 1979–1986, Sept. 2007.

[179] T.H.Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. on Audio, Speech and Language Proc.*, 2008, submitted.

[180] 3GPP2 C.S0014-0, "Enhanced variable rate codec (EVRC)," Dec. 1999.

[181] C. Morioka, A. Kurashima, and A. Takahashi, "Proposal of objective speech quality assessment for wideband IP telephony," in *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, March 2005, pp. 49–52.

[182] ITU-T Contribution COM 12-181, "Extension of the methodology for the derivation of equipment impairment factors from subjective listening-only tests towards wideband speech codecs," Intl. Telecom. Union, 2008, (Authors: S. Möller, N. Cote, V. Barriac, and A. Raake).

[183] T. H. Falk, Y. Guo, and W.-Y. Chan, "Improving robustness of image quality measurement with degradation classification and machine learning," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, Nov. 2007, pp. 503–507.