Thesis for the degree of Doctor of Philosophy

Human Perception in Speech Processing

Volodya Grancharov

KTH Electrical Engineering

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)

Stockholm 2006

Volodya Grancharov
Human Perception in Speech Processing

ISBN 91-628-6864-0
TRITA-EE 2006:016
ISSN 1653-5146

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)
SE-100 44 Stockholm, Sweden
Telephone + 46 8 790 8819

Abstract

The emergence of heterogeneous networks and the rapid increase of Voice over IP (VoIP) applications provide important opportunities for the telecommunications market. These opportunities come at the price of increased complexity in the monitoring of the quality of service (QoS) and the need for adaptation of transmission systems to the changing environmental conditions. This thesis contains three papers concerned with quality assessment and enhancement of speech communication systems in adverse environments.

In paper A, we introduce a low-complexity, non-intrusive algorithm for monitoring speech quality over the network. In the proposed algorithm, speech quality is predicted from a set of features that capture important structural information from the speech signal.

Papers B and C describe improvements in the conventional pre- and post-processing speech enhancement techniques. In paper B, we demonstrate that the causal Kalman filter implementation is in conflict with the key properties in human perception and propose solutions to the problem. In paper C, we propose adaptation of the conventional postfilter parameters to changes in the noisy conditions. A perceptually motivated distortion measure is used in the optimization of postfilter parameters. Significant improvement over nonadaptive system is obtained.

Keywords: quality assessment, non-intrusive, quality of service, postfilter, speech coding, speech enhancement, noise reduction, additive noise, multiplicative noise, tandeming, perceptually optimal processing, distortion measure, speech enhancement, optimal lag, Kalman filter, causal filter, Kalman smoother, AR model.

List of Papers

The thesis is based on the following papers:

[A] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "On causal algorithms for speech enhancement," to appear in *IEEE Transactions on Speech and Audio Processing.*, vol. 14, pp. 764-773, 2006

[B] V. Grancharov, D. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, non-intrusive speech quality assessment," *IEEE Trans. Speech, Audio Processing, special issue on Objective Quality Assessment of Speech and Audio*, submitted

[C] V. Grancharov, J. Plasberg, J. Samuelsson, and W. B. Kleijn, "Generalized postfilter for speech quality enhancement," *IEEE Trans. Speech, Audio Processing*, to be submitted

In addition to papers A-C, the following papers and patents have also been produced during the course of the PhD study:

[1] V. Grancharov, A. Georgiev, W. B. Kleijn "Sub-Pixel Registration of Noisy Images," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (ICASSP), pp. 273-276, Toulouse, France, 2006

[2] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Distortion Measures for Vector Quantization of Noisy Spectrum," *Proc. Interspeech* (ICSLP), pp. 3173-3176, Lisbon, Portugal, 2005

[3] V. Grancharov, J. Samuelsson, and W. B. Kleijn "Improved Kalman Filtering for Speech Enhancement," *Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing* (ICASSP), pp. 1109-1112, Philadelphia, USA, 2005

[4] V. Grancharov, S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Robust spectrum quantization for LP parameter enhancement", *Proc. XII European Signal Processing Conf.* (EUSIPCO), pp. 1951-1954, Vienna, Austria, 2004

[5] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Noise-dependent postfiltering," *Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing* (ICASSP), pp. 457-460, Montreal, Canada, 2004

[6] V. Grancharov and W. B. Kleijn, book chapter "Speech Quality Estimation" in *Springer Handbook of Speech Processing and Speech Communication*, J. Benesty, Y. Huang, and M. Sondhi, Eds., in preparation

[7] V. Grancharov, D. Zhao, J. Lindblom, and W. B. Kleijn, "Non-Intrusive Speech Quality Assessment with Low Computational Complexity," *Proc. Interspeech* (ICSLP), Pittsburgh, USA, submitted

[8] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Noise-dependent postfiltering," international patent filed by Nokia Corporation, 2003

[9] V. Grancharov, W. B. Kleijn, and S. Bruhn, "Low-complexity, non-intrusive speech quality assessment," provisional patent application filed by Ericsson AB, 2006

Acknowledgements

I am thankful to my supervisor Prof. Bastiaan Kleijn for sharing with me his creativity, professionalism, and dedication.

I am indebted to all previous and current members of Sound and Image Processing Lab: Anders, Arne, Barbara, David, Davor, Dora, Elisabet, Ermin, Harald, Jan, Jonas L., Jonas S., Mattias, Moo Young, Renat, Sriram. I really enjoyed working with you.

I express my gratitude to my family: my wife Nina and my daughter Mila, for their patience and understanding.

Volodya Grancharov
Stockholm, May, 2006

Contents

Acronyms

ACR:	Absolute Category Ratings
AMR:	Adaptive Multi-Rate
AMR-WB:	Adaptive Multi-Rate Wideband
ANSI:	American National Standards Institute
AR:	Autoregressive
BSD:	Bark Spectral Distortion
CELP:	Code-Excited Linear Prediction
DCR:	Degradation Category Rating
DMOS:	Degradation Mean Opinion Score
DRT:	Diagnostic Rhyme Test
EM:	Expectation Maximization
ERB:	Equivalent Rectangular Bandwidth
EVRC:	Enhanced Variable Rate Coder
GMM:	Gaussian Mixture Model
GPF:	Generalized Postfilter
IIR:	Infinite Impulse Response
ITU:	International Telecommunication Union
LCQA:	Low-Complexity Speech Quality Assessment
LP:	Linear Prediction
LSF:	Line Spectral Frequencies
MMSE:	Minimum Mean Squared Error

MNRU:	Modulated Noise Reference Unit
MOS:	Mean Opinion Score
MRT:	Modified Rhyme Test
MSE:	Mean Squared Error
MUSHRA:	Multi Stimulus Test with Hidden Reference and Anchors
PDF:	Probability Density Function
PEAQ:	Perceptual Evaluation of Audio Quality
PESQ:	Perceptual Evaluation of Speech Quality
PLP:	Perceptual Linear Prediction
PSQM:	Perceptual Speech Quality Measure
QoS:	Quality of Service
RMSE:	Root Mean Square Error
SD:	Spectral Distortion
SNR:	Signal-to-Noise Ratio
SSNR:	Segmental Signal-to-Noise Ratio
VAD:	Voice Activity Detector
VoIP:	Voice over IP

Part I

Introduction

Introduction

This thesis is about incorporating knowledge of human perception into speech quality estimation and speech quality enhancement systems. The key properties of the human perception are covered in the first part of the thesis introduction. Then the introductory part continues with a discussion of the state-of-the-art in speech quality estimation, pre-processing speech enhancement, and post-processing speech enhancement. The main body of the thesis consists of three articles that present the contributions of the author to the problems discussed in the introduction.

1 Introduction to Human Perception

Sound is a longitudinal pressure wave consisting of compressions and rarefactions of air molecules. Compressions are zones where air molecules have been forced into a tighter configuration by the application of energy, and rarefactions are zones where air molecules are less tightly packed, see Fig. 1.

As sound travels as pressure waves through the air, it is collected by the *pinna* of the *outer ear*, Fig. 2. The outer ear includes also the *auditory canal* that ends at the *ear drum*. Through the auditory canal, which is air-filled, the sound is carried to the ear drum located in the *middle ear*. The auditory canal filters the sound, giving a resonance at approximately 5 kHz. The middle ear space is connected to the back of the throat by the *eustachian tube*. The eustachian tube is normally closed, but opens when we swallow, equalizing the middle ear pressure with the external air pressure. The middle ear mechanically conveys the sound pressure to the ear drum, exciting the fluid in the *cochlea*. The mechanical middle ear system not only conveys, but amplifies the pressure forced on the fluid. The main purpose of the cochlea is to transfer the pressure changes of the fluid to neural firings in the *auditory nerve*.

The process of transduction (transforming mechanical vibrations into electrical signals) is performed by specialized sensory cells within the cochlea. There are approximately 3 500 inner hair cells and 11 000 outer hair

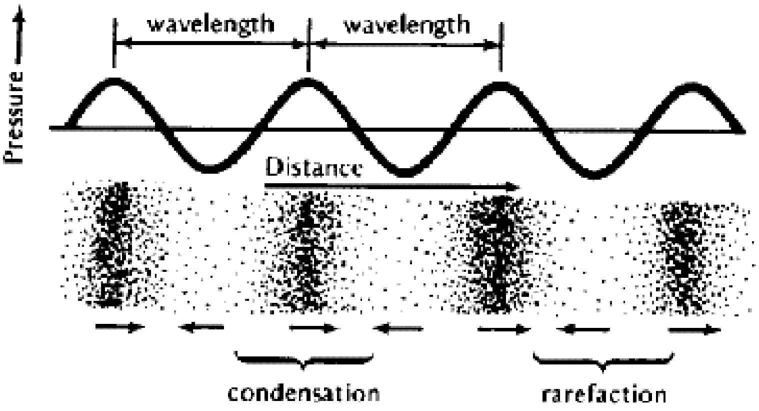

Figure 1: A longitudinal pressure wave.

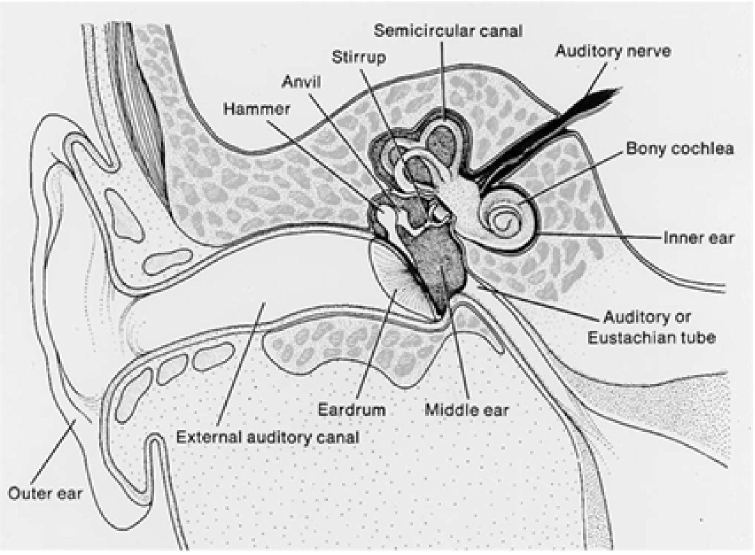

Figure 2: The human peripheral auditory system consists of three parts: the outer, middle, and inner ear. The function of the outer ear is to collect the signal. In the middle and inner ear the acoustical waves are transformed into nerve impulses, transmitted to the brain.

cells. These hair cells connect to approximately 24 000 nerve fibers. The rocking of the stirrup in the oval window shakes the fluid within the cochlear causing movement of the hair cells. The cochlea acts as if it were made up of overlapping filters having bandwidths equal to the critical bandwidth. The filters closest to the cochlear base respond to the higher frequencies, and those closest to its apex respond to the lower frequencies.

The outlined peripheral auditory organ (ear) is the first major component of the auditory perception system, shown in Fig. 3. It processes an acoustic pressure signal by first transforming it into a mechanical vibration pattern on the basilar membrane, and then representing the pattern by a series of pulses to be transmitted by the auditory nerve. The second major component of the auditory perception system is the auditory nervous system (brain), where cognitive processing is performed.

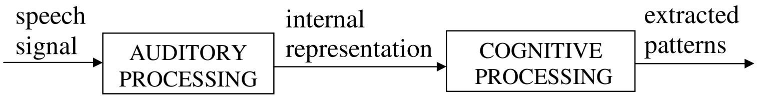

Figure 3: Low- and high-level processing steps in the sound perception mechanism.

The way in which the brain processes extracted patterns is largely unknown. Many studies have shown how humans perceive tones and bands of noise [1], [2]. Based on that knowledge, many auditory models that simulate the functionality of the human ear, have been created [1–4].

It is well known that the ear's frequency resolution is not uniform on the Hertz scale. The peripheral auditory system contains a bank of bandpass filters with overlapping passbands. The bandwidth of each auditory filter is called the critical bandwidth. Commonly used quantitative description of the critical bandwidth is the Equivalent Rectangular Bandwidth (ERB). Each ERB band corresponds to a width of approximately 0.9 mm on the basilar membrane. The conversion from Hertz f to ERB scale is given by:

$$ERB(f) = 0.108\ f + 24.7. \tag{1}$$

Other perceptually based scales are the Bark and Mel scales. The conversion from Hertz to Bark b frequency scale is defined as:

$$b(f) = 6\ \sinh^{-1}\left(\frac{f}{600}\right). \tag{2}$$

A third perceptually motivated scale is the Mel frequency scale, which is linear below 1 kHz and and logarithmic above that frequency:

$$m(f) = 1127\ \ln\left(1 + \frac{f}{700}\right). \tag{3}$$

A well-established fact is that *perceived loudness* (a subjective measure of sound intensity) is related to signal intensity in a complex, nonlinear way. A logarithmic function is typically used as a rough approximation to convert the signal intensity to perceived loudness [5].

An important property of human auditory system is the non-uniform *equal loudness* perception of tones of varying frequencies. In general, tones of differing pitch have different inherent perceived loudness. The sensitivity of the ear varies with frequency. The ear's sensitivity is not only a function of frequency, but of absolute hearing thresholds as well, as shown in Fig. 4.

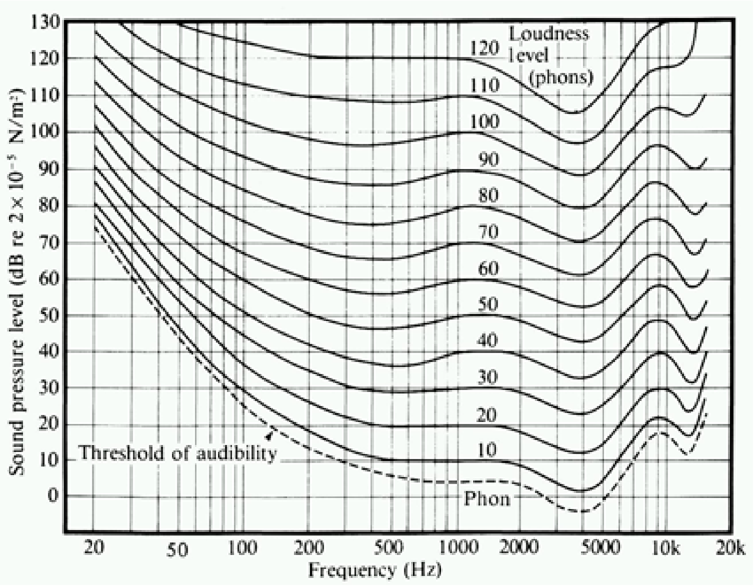

Figure 4: Equal loudness contour diagram.

Many studies have demonstrated time- and frequency-masking effects. Masking is defined as the increase of the threshold of audibility of one sound (*maskee*) in the presence of another sound (*masker*). The masking may occur simultaneously in time (frequency masking), as illustrated in Fig. 6. Another form of masking is non-simultaneous (forward or backward time masking), shown in Fig. 5.

Despite of the significant progress in the area of psychoacoustics, there are still open questions to be answered, particularly with respect to complex signals. Most of the psychoacoustical experiments are performed with simple sounds. However, speech (which is the focus of this thesis) is a complex and dynamic signal, which is not always perceived as a superposition of its

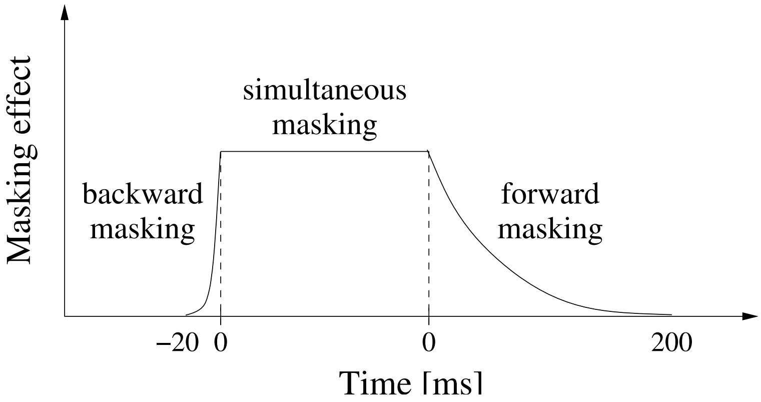

Figure 5: Non-simultaneous masking occurs before and after the masker.

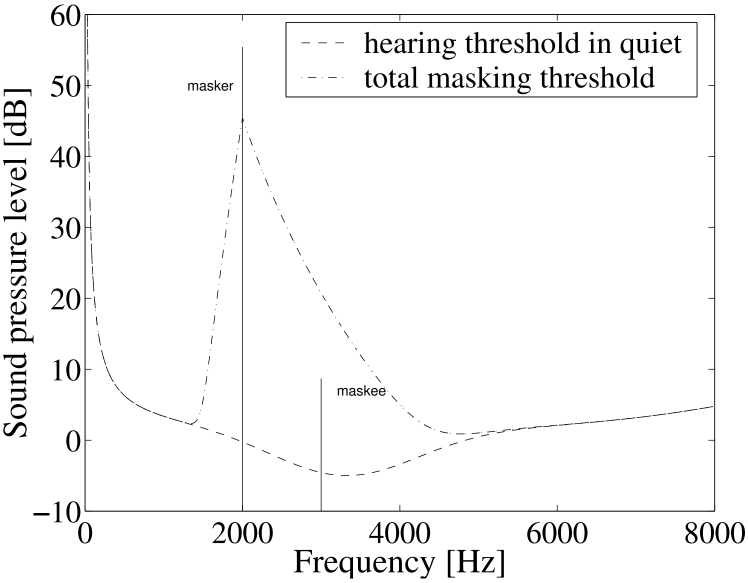

Figure 6: Simultaneous masking occurs when a strong tone makes the nearby tone inaudible.

basic components. The perception of a complex signal, such as speech, is not well understood. Some evidence of the importance of the dynamics in the speech signal is presented in [6–8].

Incorporation of the knowledge of human auditory processing in state-of-the-art speech enhancement systems is the essence of papers B and C, presented in this thesis. In the past a number of psychoacoustical concepts have been integrated successfully into speech and audio coding [9–17].

In paper B we study the perceptual differences between the causal and non-causal implementations of the widely used linear mean squared error filters. After demonstrating that the causal implementation is in conflict with human perception, we propose improvements on the existing systems.

The focus of paper C is on the adaptation of the commonly used speech coding postfilter to changes in environmental conditions. The proposed adaptation is based on an advanced psychoacoustical model. The postfilter structure itself is based on the masking properties of the human auditory system, and its parameters are set based on listening tests.

The discussion so far has been concerned with the low-level processing step of the human auditory system, where the speech waveform is transformed into a nerve excitation. The importance of the high-level processing performed by the brain is demonstrated in paper A. We hypothesize that at the high-level processing step, performed by the brain, structural information is extracted from the signal and compared with already stored patterns. This was confirmed by the test results of the performed simulations. Furthermore, the proposed speech quality assessment measure demonstrated higher accuracy than the current state-of-the-art.

2 Speech Quality Estimation in Telecommunication Systems

Speech communication systems, and especially VoIP systems, can suffer from significant call quality degradation, caused by noise, echo, etc. [18]. Internet protocol (IP) networks guarantee neither sufficient bandwidth for the voice traffic, nor a constant, acceptable delay. Dropped packets and varying delays introduce distortions not found in traditional telephony. In addition, if a low bit-rate codec is used in VoIP to achieve a high compression ratio, the original waveform can be significantly distorted. All these factors can affect psychological parameters like *intelligibility*, *naturalness*, and *loudness* that determine the overall speech quality. The influence of physical network parameters on psychological quality parameters is summarized in Table 1.

There are two broad classes of speech quality metrics: subjective and objective. Subjective measures involve humans listening to a live or recorded conversation and assigning a rating to it. Objective measures are computer algorithms designed to estimate quality degradation in the signal. Speech quality is a complex psycho-acoustic phenomenon within the process of human perception. As such, it is necessarily subjective, even different people interpret speech quality differently. However, the objective measures are

Table 1: Different physiological characteristics of speech quality and their dominant dependencies on physical network characteristics. *Intelligibility* measures the quality of the perception of the meaning or information content of what the talker has said. *Naturalness* is the degree of fidelity to the talker's voice. *Loudness* is the absolute loudness level at the listener's side. The symbol "+" denotes dependency on the parameter.

	Psychological Parameters			
Physical Parameters	Intelligibility	Naturalness	Loudness	*Quality*
Signal Level	+	+	+	+
Noise	+			+
Freq. Response	+	+	+	+
Distortion	+	+		+
Delay	+			+
Echo	+			+
Packet Loss	+			+

Table 2: Comparison of Subjective and Objective Methods for Quality Estimation. The symbol "+" is used to denote that the method is advantageous over the other method, denoted by "-".

	Subjective Measures	Objective Measures
Cost	-	+
Reproducibility	-	+
Automation	-	+
Unforeseen Impairments	+	-

widely used since they have several critical advantages over the subjective measures, see Table 2.

2.1 Subjective Measures

In subjective tests, human participants assess the performance of a system in accordance with opinion scale [19], [20]. Two general categories of subjective quality measures are *conversational quality* measures and *listening* quality measures. *Conversational* quality refers to how listeners rate their ability to converse during the call (which includes listening quality). In conversational tests, a pool of listeners are placed into interactive communication scenarios, and asked to complete a task over the phone. By evaluating the efficacy of the performance of the task, the listeners provide a quality measure for effects like delay, echo, and loudness. *Listening quality* refers to how listener rate what they "hear" during the call. Listening quality ignores effects such as echoes at the talker side or transmission delays.

In an Absolute Category Ratings (ACR) test, a pool of listeners rate a series of audio files using a five level impairment scale. After obtaining individual scores, the mean opinion for each audio file is calculated. To achieve reliable results, test are performed with a large pool of listeners and under controlled conditions. Mean Opinion Score (MOS) is the most widely used method to evaluate the overall speech quality. MOS is a five level scale from "Bad" do "Excellent", as shown in Table 3.

Table 3: Table of grades in the MOS scale.

Bad	1
Poor	2
Fair	3
Good	4
Excellent	5

In Degradation Category Rating (DCR) tests, listeners hear the reference and the test signals sequentially, and are asked to compare them. Degradation MOS (DMOS) is an impairment grading scale to measure how the different distortion in speech are perceived, see Table 4.

Table 4: Table of grades in the DMOS scale. Listeners are asked to describe degradation in the signal.

Very annoying	1
Annoying	2
Slightly annoying	3
Audible, but not annoying	4
Inaudible	5

A variation on the DCR test is a Comparison Category Rating (CCR) test. Listeners identify the quality of the second stimulus relative to the first one on the scale presented in Table. 5.

DCR tests are more common in audio quality assessment [21, 22], while speech coding systems are typically assessed by an ACR test. One example of a DCR test is a MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) [21], a method for the subjective assessment of intermediate quality level of coding systems. MUSHRA is a double-blind multi-stimulus test method with a hidden reference and hidden anchors. In this test, the subjects are required to score the stimuli according to the continuous quality scale from 0 to 100. The listener records his/her assessment of the quality

Table 5: Table of grades in the CCR test. Listeners grade the perceived quality of a speech signal in relation to a reference speech signal.

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

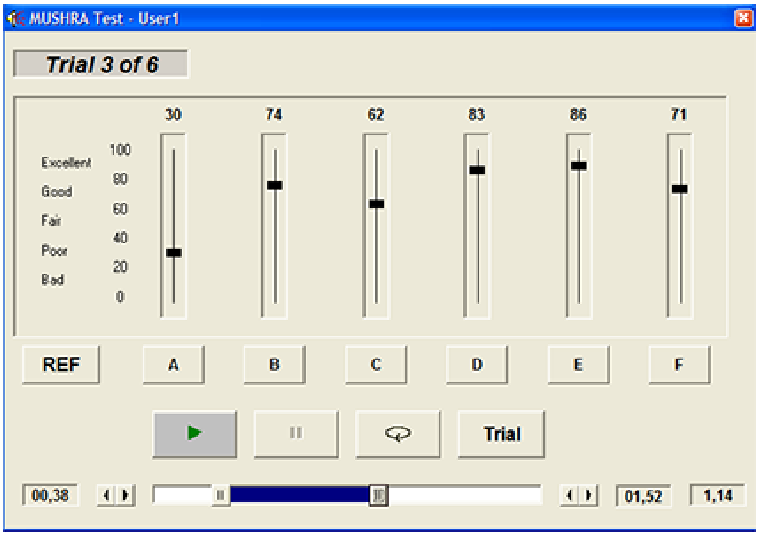

Figure 7: Graphical user interface for the MUSHRA test. The test subject can compare the files under test (buttons A-F) with the original signal (button REF).

with the use of sliders on an electronic display, see Fig. 7.

A classification of the most popular ACR and DCR tests, standardized by the ITU, is presented in Fig. 8. Major conceptual differences between the two tests are: 1) in ACR even an original signal can receive low grade, since listeners compare with their internal model of "clean speech", 2) DCR tests provide a quality scale of higher resolution, due to comparison of the distorted signal with one or more reference/anchor signals.

A procedure that is not so commonly used nowadays is Diagnostic Acceptability Measure (DAM) [23]. It provides more systematic feedback and evaluates speech quality on 16 scales. In contrast to most other measures,

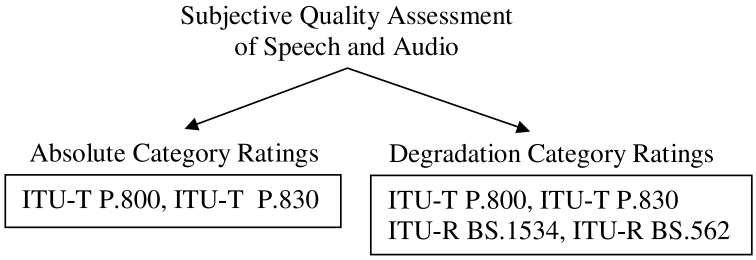
Subjective Quality Assessment of Speech and Audio
Absolute Category Ratings
ITU-T P.800, ITU-T P.830
Degradation Category Ratings
ITU-T P.800, ITU-T P.830
ITU-R BS.1534, ITU-R BS.562

Figure 8: The two major types of subjective quality assessment methods and related ITU standards and recommendations.

trained listeners are used in the DAM test. A weighted average of all scales forms a composite measure that describes the condition under test.

An example of an intelligibility test is the Diagnostic Rhyme Test (DRT), which uses a set of isolated words to test for consonant intelligibility in the initial position. The test consists of 96 word pairs that differ by a single acoustic feature in the initial consonant. The Modified Rhyme Test (MRT) [24] is an extension to the DRT. It tests for both initial and final consonants. A set of six words is played one at a time and the listener marks which word he/she thinks he/she hears.

Reference conditions (well defined conditions) of processed speech are commonly used in listening tests. The most popular one is the Modulated Noise Reference Unit (MNRU) [25]. The MNRU is a reference condition that adds amplitude modulated noise to a speech signal. The main reason to introduce MNRU conditions is that they can provide a spread in quality level, which increases the accuracy of the human ratings.

2.2 Objective Measures

Subjective listening or conversational tests can be used to gather first-hand evidence about perceived speech quality, but such tests are often expensive, time-consuming, and labor-intensive. Objective quality algorithms can be used instead, but they have to be properly "calibrated" to the output of subjective quality tests.

Typically, the accuracy of an objective metric is determined by its correlation with MOS scores for a set of data. The estimation performance is assessed using the correlation coefficient R and the root-mean-square error (RMSE) ε, between the predicted quality $\hat{Q}$ and the measured subjective quality Q. The RMSE is given by

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{N}(Q_i - \hat{Q}_i)^2}{N}}, \tag{4}$$

and the correlation coefficient is defined as

$$R = \frac{\sum_{i=1}^{N}(\hat{Q}_i - \mu_{\hat{Q}})(Q_i - \mu_Q)}{\sqrt{\sum_{i=1}^{N}(\hat{Q}_i - \mu_{\hat{Q}})^2}\sqrt{\sum_{i=1}^{N}(Q_i - \mu_Q)^2}}, \tag{5}$$

where μ_Q and $\mu_{\hat{Q}}$ are the mean values of the introduced variables and N is the number of MOS labeled utterances used in evaluation. The evaluation is typically done over a large multi-language database that contains a wide range of distortions, e.g., [26].

Some objective quality measures are designed to estimate the listening subjective quality, while others estimate the conversational subjective quality. Alternatively, the classification of objective quality measures can be based on the type of input information they require: intrusive quality measures require access to both the original and distorted speech signal, while the non-intrusive measures base their estimate only on the distorted signal. A general classification of objective quality measures and the corresponding ITU standards is presented in Fig. 9.

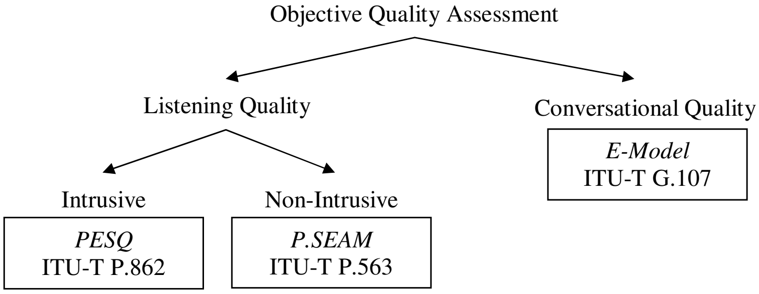

Figure 9: Classification of objective quality assessment methods and related ITU standards.

Intrusive Listening Quality Measures

Historically, most objective quality measures are designed to estimate subjective *listening quality* in an *intrusive* manner. The simplest and most common quality assessment measures are SNR and SSNR. The overall SNR distortion measure between an original $\mathbf{s}$ and distorted $\mathbf{y}$ speech vectors is calculated as:

$$\mathrm{d}_{SNR}(\mathbf{s}, \mathbf{y}) = 10 \log_{10}\left(\frac{\mathbf{s}^T\mathbf{s}}{\mathbf{e}^T\mathbf{e}}\right), \tag{6}$$

where $\mathbf{e} = \mathbf{s}$ - $\mathbf{y}$. The vector dimension is sufficient to contain the entire utterance.

The SSNR is calculated by splitting the two vectors into smaller blocks and calculating a SNR value for each of these blocks. The final SSNR value is obtained by averaging the per-block SNR values:

$$\mathrm{d}_{SSNR}(\mathbf{s},\mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N} 10\log_{10}\left(\frac{\mathbf{s}_n^T\mathbf{s}_n}{\mathbf{e}_n^T\mathbf{e}_n}\right), \tag{7}$$

where N is the total block number, n is the block index, and the per-block error vector is defined as $\mathbf{e}_n = \mathbf{s}_n - \mathbf{y}_n$. A typical block length is 5 ms.

SNR and SSNR are simple to implement, have straightforward interpretations, and can provide indications of perceived speech quality for a specific waveform-preserving speech systems [27]. Unfortunately, when used to evaluate coding and transmission systems in a more general context SNR and SSNR show little correlation to perceived speech quality.

Frequency-domain measures are known to be significantly better correlated with human perception, but still relatively simple to implement. One of their critical advantages is that they are less sensitive to signal misalignment. Perhaps the most popular frequency domain measure is the gain-normalized SD, which is widely accepted as a quality measure of coded speech spectra. It evaluates the similarity of two autoregressive envelopes:

$$\mathrm{d}_{SD}(\mathbf{s},\mathbf{y}) = \frac{1}{N}\sqrt{\sum_{n=1}^{N}\int_{-\pi}^{\pi}\left(10\log_{10}\left(\frac{P_{\mathbf{s}}(\omega,n)}{P_{\mathbf{y}}(\omega,n)}\right)\right)^2\frac{d\omega}{2\pi}}, \tag{8}$$

where N is the total number of frames, $P_{\mathbf{s}}(\omega,n)$ and $P_{\mathbf{y}}(\omega,n)$ are the autoregressive spectra of the clean and processed signal. Other popular frequency domain measures include the Itakura-Saito, Log-Likelihood, and Log-Area-Ratio measures.

During the last two decades the researchers have moved their focus to the class of perceptual domain measures. These measures are based on models of human auditory perception. The Bark Spectral Distortion (BSD) is one of the first objective measures based entirely on a model of human perception [28]. It calculates the averaged Euclidean distance between the original and distorted speech signals in the Bark domain.

Perceptual Speech Quality (PSQM) [29] is a perceptually motivated speech quality assessment algorithm, designed to assess the performance of speech codecs and impairments encountered in networks. Since the accuracy of PSQM was not sufficient, the most successful measures, evaluated by the ITU in the 1990s, were combined into an improved model Perceptual Evaluation of Speech Quality (PESQ), which was accepted as ITU recommendation in 2001 [30]. Like PSQM, PESQ is intended to be used for measuring quality of narrowband telephone signals. PESQ is certified to provide speech quality estimate in the following environments: speech codecs, transmission channel errors, speech input level at the codec, noise added by the system, time warping, packet loss, and time clipping. The current research focus is on the development of a wide-band extension for PESQ [31].

Significant standardization efforts have been made in the area of objective audio quality assessment. These efforts resulted in the development of the Perceptual Evaluation of Audio Quality (PEAQ) measure [32], which is the ITU standard for audio quality assessment.

The PSQM, PESQ, and PEAQ algorithms for quality estimation are based on the following algorithmic blocks: 1) the signals are processed by a filter that simulates the frequency response of a typical telephone headset, 2) a Hoth noise is injected to model a typical listening environment, 3) an intensity warping is performed, to model the relationship between signal power and perceived loudness, 4) a loudness scaling is performed to equalize the momentary compressed loudness of the two signals, and 5) the distance between the transformed signals is calculated and mapped to an estimate of MOS value. The general scheme of the perceptually motivated distortion measures, is presented in Fig. 10.

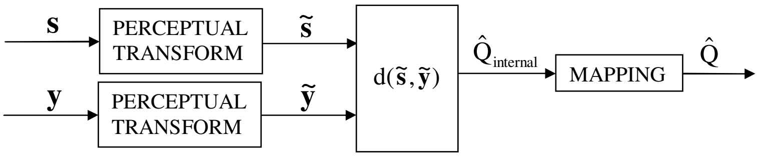

Figure 10: The distance between signals is calculated after applying a perceptual transform.

The final part of the human judgement process entails cognitive processing in the brain, where compact features are extracted from auditory excitations. It is easy to notice that the forementioned objective quality assessment algorithms incorporate knowledge of the low-level auditory processing, but neglect the high-level cognitive processing, performed by the brain. One exception is the Measuring Normalizing Blocks (MNB) algorithm [33], [34], which utilizes a relatively simple perceptual transform, but a sophisticated error pooling system. Another example can be found in [35], where the authors recognize the importance of the high-level cognitive process and apply a statistical data mining approach. In the approach of [35], a large pool of candidate features is created and the ones that lead to the most accurate prediction of perceived quality are selected. In Fig. 11 the desired desired (which is not realizable with the current knowledge of high-level cognitive processes, as performed by the human brain) is illustrated.

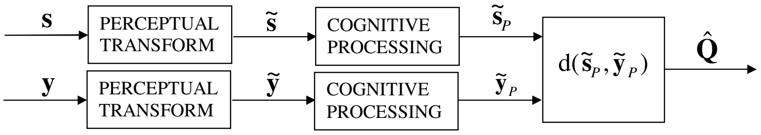

Figure 11: Desired scheme of perceptually motivated speech quality assessment measure.

The differences between Fig. 10 and Fig. 11 demonstrate the weakness of the majority of existing perceptually motivated speech quality assessment measures. These algorithms exploit the knowledge of the human auditory system to weight more the error signal in regions where it is more audible. However, more audible does not necessarily mean more objectionable, since the latter is dependent of the a-priori information in the human brain. There is no guarantee that less audible parts of the signal may not be of higher importance for the pattern extraction and comparison process performed by the human brain, after the signal has been perceptually transformed.

Non-Intrusive Listening Quality Measures

In many applications requiring speech quality assessment, the original speech signal may not be available, or it may be difficult to align it to the processed speech signal. In such cases, an attractive alternative approach is to predict the speech quality from the processed signal only. Such a type of quality assessment is important in monitoring of communication systems, such as wireless communications and VoIP. An objective measure for non-intrusive speech quality assessment based on the temporal envelope representation of speech can be found in [36]. A different approach to non-intrusive quality assessment is presented in [37], where the authors model the limitations of the human vocal tract and estimate the level of speech distortion from the parameters that violate the resulting constraints.

The majority of non-intrusive quality assessment algorithms perform a similar perceptual transform on the input signal, but offer a large variety of mapping schemes [38–41], such as Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Neural Networks, etc. The ITU standard of non-intrusive speech quality assessment can be found in [42]

A non-intrusive speech quality assessment system, based on a speech spectrogram, is presented in [43]. An interesting concept in this approach is that accurate estimation of speech quality is achieved without a perceptual transform of the signal. Similar concepts can be found also in recent advances in image quality assessment, e.g. [44].

Objective Measures for Assessment of Conversational Quality

The objective measure that provides an estimate of the conversational subjective quality is the E-Model [45]. In contrast to the previously described schemes, the E-Model is a purely parametric model. It is a transmission rating model that monitors many different parameters and combines their values into an end-performance factor. The E-Model was originally used as a network planning tool, but it has gained a wider acceptance and nowadays is used non-intrusively over the network as a passive monitoring tool.

The objective of the E-model is to determine a transmission quality rating, i.e., the ”R” factor, with range typically between 0 and 120. The ”R” factor can be converted to estimated listening and conversational quality MOS scores. The E-model does not compare the original and received signals directly. Instead, it uses the sum of equipment impairment factors, each one quantifying the distortion due to a particular factor. Impairment factors include the type of speech codec, echo, averaged packet delay, packet delay variation, and the fraction of packets dropped. As an example, let us consider a system with distortion due to the codec I_{codec}, averaged one-way delay I_{delay}, packet delay variation I_{dv}, and packet loss $I_{packetloss}$. Then, the transmission quality factor can be calculated as:

$$R = R_0 - I_{codec} - I_{delay} - I_{dv} - I_{packetloss}, \tag{9}$$

where R_0 is the highest possible rating for this system. The broader scope of conversational quality assessment, as compared to listening quality assessment, is illustrated in Fig. 12. Note that both P.SEAM and E-Model are non-intrusive, i.e., they do not require the original signal(s).

The discussed measures of listening and conversational quality are designed to predict the speech quality from the simultaneous effect of large number of distortions. An objective quality assessment measure can also be designed to operate in a particular environment only (e.g., specific speech coding standard). These constraints can significantly improve the accuracy of the system and reduce complexity and memory requirements [46].

3 Pre-Processing Speech Enhancement Techniques

Historically, pre-processor single-channel speech enhancement algorithms have been considered in the context of robust speech coding, see Fig. 2. These algorithms are designed to operate in an environment where only the noisy signal is available [47], and both facilitate the operation of the speech codec and improve the perceived sound quality at the end user.

In a single-channel application, the noise suppression algorithm requires an additional module for the estimation of the noise and clean speech statis-

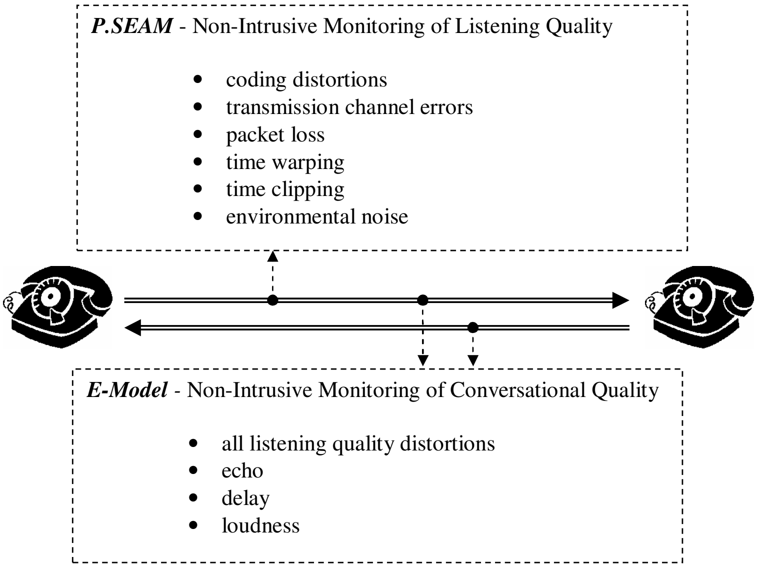
P.SEAM - Non-Intrusive Monitoring of Listening Quality
coding distortions
transmission channel errors
packet loss
time warping
time clipping
environmental noise
E-Model - Non-Intrusive Monitoring of Conversational Quality
all listening quality distortions
echo
delay
loudness

Figure 12: Non-intrusive monitoring of listening and conversational quality over the network.

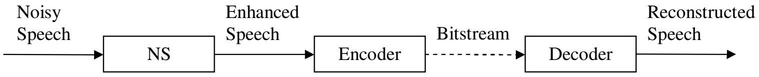

Figure 13: Configuration of noise suppression (NS) as a speech enhancement pre-processor for speech codec.

tics. Some of the most commonly used voice activity detectors (VAD) and soft-decision methods can be found in [48–52]. The underlying idea in all these algorithms is that the noise statistics can be estimated from the signal segments, either in the time or in the frequency domain, where the speech energy is either low, or the speech signal is not present at all.

The classical noise suppression scheme is based on the idea of spectral subtraction [53]. It is widely used nowadays, mainly because of its simplicity. Spectral subtraction schemes are based on direct estimation of the short-time spectral magnitude of clean speech. A drawback of this algorithm is the *musical noise* effect [54], [55]. Musical noise consists of tones with the same duration as the window length of algorithm and with a different set of frequencies for each frame. Musical noise is a result of variability in the power spectrum.

In an attempt to improve on the perceptual performance, a generalization of spectral subtraction was proposed, in the form of nonlinear spectral subtraction [56]. A theoretically motivated approach to improve on speech and noise parameter estimation is proposed in [57].

Speech enhancement can be based on a signal subspace methods [58], [59], or wavelet based methods [60–62]. In signal subspace methods, speech distortion is minimized, subject to a constraint on a residual noise level. In practice, both wavelet and subspace methods achieve noise reduction through thresholding.

The use of models for speech and/or noise improve the performance of speech enhancement systems. Different models for speech and noise have been investigated: the sinusoidal model was used in [63], the autoregressive model in [64], [65], [66]. More advanced modelling, based on HMM, is used to capture speech dynamics in [67], [68].

A-priori information may be incorporated in the noise suppression algorithms not only through the type of the model, but also in the form of model parameters. Recent advances in noise suppression algorithms exploiting a-priori speech and noise information, in the form of parameters of AR processes, can be found in [69] and [70].

Due to the constant interest from the speech coding industry many attempts have been made for standardization of noise suppression algorithms. Examples of standardized algorithms can be found in [71–73]. Because of the complexity of the problem, none of the candidate algorithms passed the minimum requirements, in the recent standardization effort [74]. Current state-of-the-art public algorithms are described in [75–77].

In the following, we consider only noise suppression algorithms designed to improve the quality of the perceived speech signal. For completeness mention that noise suppression pre-processors are also used in the context of robust speech and speaker recognition, or in noise suppression systems optimized for the performance of the speech codec parameters [78].

3.1 Linear Minimum Mean-Squared Error Filters

Let us consider the problem of observing a speech signal in the presence of additive noise:

$$y_k = s_k + v_k. \tag{10}$$

With y_k, s_k and v_k we denote discrete-time samples of noisy speech, clean speech and noise, respectively. We assume that the signals are random processes and that speech and noise are uncorrelated and zero-mean.

Let $\mathbf{s} = [s_L \dots s_1]$ denote a segment of length L of the clean speech signal, and the noisy observation $\mathbf{y}$ is defined analogously. Let us consider the optimal estimator of $\mathbf{s}$, given only the statistically related noisy observations, in the mean-squared error sense. That is, we seek the estimate $\hat{\mathbf{s}}$ that minimizes

$$E\{(\mathbf{s}-\hat{\mathbf{s}})(\mathbf{s}-\hat{\mathbf{s}})^T\}. \tag{11}$$

We search for the optimal estimator as an arbitrary function of the observation $\mathbf{y}$, say $\hat{\mathbf{s}} = g(\mathbf{y})$. It is well known that the solution to (11), the optimal minimum mean-squared estimator of a random variable $\mathbf{s}$ given the value of another random variable $\mathbf{y}$, is given by the conditional expectation, e.g., [79]:

$$\hat{\mathbf{s}} = E\{\mathbf{s}|\mathbf{y}\} \doteq \int_{-\infty}^{+\infty} \mathbf{s} f(\mathbf{s}|\mathbf{y}) d\mathbf{s}, \tag{12}$$

where $f(\mathbf{s}|\mathbf{y})$ is the conditional pdf of $\mathbf{s}$, given $\mathbf{y}$.

In this thesis, we consider the problem of finding a *linear* minimum mean-squared estimator and study applications of smoother and filter in speech enhancement. We note that for Gaussian variables, the linear estimator is the optimal estimator, e.g., [80]. Thus, an equivalent starting point would have been the assumption of Gaussianity for our signals. In the case of a linear *filter*, the estimate is based only on the past and current observations:

$$\hat{s}_k^F = E\{s_k|y_k, y_{k-1}, \dots, y_1, y_0\} \tag{13}$$

The *smoother* is based on a certain amount of future noisy observations, in addition to the past and present observations:

$$\hat{s}_k^S = E\{s_k|y_{k+M}, y_{k+M-1}, \dots, y_k, \dots, y_1, y_0\}. \tag{14}$$

A consistent theory that deals with the data-dependent linear MMSE filters was first formulated by Norbert Wiener [81]. The name of Norbert Wiener is typically associated with the non-causal formulation of the optimal linear mean squared-error estimator of s_k given all the observations $\{y_m\}_{m=-\infty}^{+\infty}$:

$$\hat{s}_k = \sum_{m=-\infty}^{+\infty} h_m y(k-m). \tag{15}$$

The frequency response of the IIR Wiener filter, which is the solution to the above posed problem is given by

$$H(e^{j\omega}) = \frac{P_{sy}(e^{j\omega})}{P_y(e^{j\omega})}, \tag{16}$$

where $P_y(e^{j\omega})$ is the power spectrum of the noisy signal, and $P_{sy}(e^{j\omega})$ is the cross-power spectrum. If the noise and the signal are uncorrelated, we have the relation $P_{sy}(e^{j\omega}) = P_s(e^{j\omega})$, where $P_s(e^{j\omega})$ is the power spectrum of the clean signal. This holds true for the application of speech observed in additive background noise. A difficulty with the Wiener filter is that the

$P_s(e^{j\omega})$ is not known and must be estimated by subtracting the estimated noise power spectrum from the noisy-speech power spectrum.

In some applications, it is desirable to minimize or avoid system delay. In such a case, the estimate is to be based only on the current and past observations:

$$\hat{s}_k = \sum_{m=0}^{+\infty} h_m y(k-m). \tag{17}$$

This problem turns out to be considerably more complex. A spectral factorization has to be performed first, $P_y(z) = \sigma_0^2 Q(z) Q(1/z)$, and then the causal IIR Wiener filter can be found [82], [83]:

$$H(e^{j\omega}) = \frac{1}{\sigma_0^2 Q(z)} \left[\frac{P_{sy}(z)}{Q(1/z)} \right]_+ . \tag{18}$$

The operator $[\cdot]_+$ yields the "causal (positive-time) part". The difficulty of performing spectral factorization is the main reason for not using the optimal causal Wiener filter in speech enhancement applications.

The problem of spectral factorization and, therefore, the causal filter implementation, is overcome by the Kalman filter theory. It offers a method to recursively obtain the estimates (13) and (14). This theory has a number of advantages over the previously discussed Wiener filters: 1) Kalman filters can be used with non-stationary signals, 2) Kalman filters can be extended easily to the vector case, 3) Kalman filters require only a finite number of past observations.

The above listed properties make the Kalman filters attractive for speech enhancement applications. Kalman filtering techniques were first applied to speech enhancement for white-noise case in [84], and later extended to colored noise [85]. Most of the studies, concerned with the application of Kalman filtering in single-channel speech enhancement, focus on parameter estimation schemes, e.g., [86], [87], [88], [89]. Different iterative schemes for joint parameter and signal estimation are proposed in [85], [90], and [91].

In the following, we shall introduce the notation needed for the definition of the Kalman filtering recursion in the context of the speech enhancement. As is standard practice, we model the speech as an autoregressive process:

$$s_k = \sum_{j=1}^{p} a_j s_{k-j} + w_k, \tag{19}$$

where w_k is a white noise excitation process and the speech model order is typically set to $p = 10$ for 8 kHz sampled speech. Equations (1) and (2) can be represented in state space form:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{F}\,\mathbf{x}_k + \mathbf{G}\,w_k \\ y_k &= \mathbf{H}^T \mathbf{x}_k + v_k, \end{aligned} \tag{20}$$

where $\mathbf{x}_k = [s_k \; s_{k-1} \; \ldots \; s_{k-p+1}]^T$ is a p-dimensional state vector, and $\mathbf{G} = \mathbf{H} = [1 \; 0 \; \ldots \; 0]^T_{p\times 1}$. The state transition matrix is given by:

$$\mathbf{F} = \begin{pmatrix} a_1 & a_2 & \cdots & a_{p-1} & a_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}_{p\times p} . \tag{21}$$

The presented speech model is not unique. A speech model that is more closely related to the speech production mechanism is proposed in [92]. An extension based on the ARMA model is discussed in [93]. However, for the sake of simplicity in this presentation we follow the model defined by (2 - 3).

Assuming that the signal and noise parameters are known, the optimal minimum mean-square linear state estimate is obtained using the Kalman filter equations [79]:

$$\begin{aligned} \mathbf{P}_{k|k-1} &= \mathbf{F}\mathbf{P}_{k-1|k-1}\mathbf{F}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T \\ \mathbf{K}_k &= \mathbf{P}_{k|k-1}\mathbf{H}(\mathbf{R} + \mathbf{H}^T\mathbf{P}_{k|k-1}\mathbf{H})^{-1} \\ \mathbf{P}_{k|k} &= [\mathbf{I} - \mathbf{K}_k\mathbf{H}^T]\mathbf{P}_{k|k-1} \\ \hat{\mathbf{x}}_{k|k-1} &= \mathbf{F}\hat{\mathbf{x}}_{k-1|k-1} \\ \hat{\mathbf{x}}_{k|k} &= \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(y_k - \mathbf{H}^T\hat{\mathbf{x}}_{k|k-1}), \end{aligned} \tag{22}$$

where $\mathbf{K}_k$ is the Kalman gain and $\hat{\mathbf{x}}_{k|k}$ and $\hat{\mathbf{x}}_{k|k-1}$ are the filtered and predicted estimate of the state. The prediction-error covariance is given by $\mathbf{P}_{k|k-1} = E\{(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})^T\}$ and $\mathbf{P}_{k|k} = E\{(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})^T\}$ is the filtering-error covariance. The measurement and driving noise variances are given by $\mathbf{R} = \sigma_v^2$ and $\mathbf{Q} = \sigma_w^2$. At each time instant the speech sample estimate can be obtained by $\hat{s}_k = \mathbf{H}^T\hat{\mathbf{x}}_{k|k}$.

It is relevant to discuss the differences between time-varying and time-invariant [79] system. The Kalman filter can also be implemented in a steady-state mode, which has computational advantages. For the stationary case it is easy to note that the error covariance $\mathbf{P}_{k|k-1}$ and the Kalman gain $\mathbf{K}_k$ are dependent only on the data statistics, but not on the actual observations $\{y_k\}$ and, therefore, can be pre-computed before the filter is actually started. The error covariance can be found as a solution of the steady-state discrete-time Riccati equation, e.g. [80]:

$$\mathbf{P} = \mathbf{F}\mathbf{P}\mathbf{F}^T - \mathbf{F}\mathbf{P}\mathbf{H}^T(\mathbf{R} + \mathbf{H}^T\mathbf{P}\mathbf{H})^{-1}\mathbf{H}\mathbf{P}\mathbf{F}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T. \tag{23}$$

Let $\bar{\mathbf{P}}$ be the positive definite solution of (23), then the stationary filter gain can be found as:

$$\mathbf{K} = \bar{\mathbf{P}}\mathbf{H}^T(\mathbf{R} + \mathbf{H}^T\bar{\mathbf{P}}\mathbf{H})^{-1}. \tag{24}$$

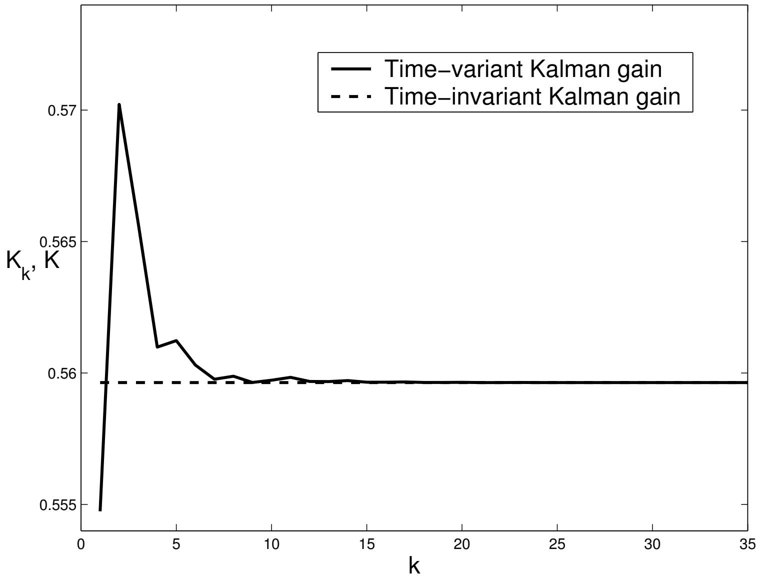

Figure 14: Stationary and time-varying Kalman gain for a representative voiced speech segment.

A simple way to find the solution of (23) is to iterate and use the fact that $\lim_{k\to\infty} \mathbf{P}_{k|k-1} = \bar{\mathbf{P}}$. After the stationary Kalman gain is obtained, the Kalman algorithm reduces to:

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}(y_k - \mathbf{H}^T \hat{\mathbf{x}}_{k|k-1}). \tag{25}$$

The use of the time-invariant Kalman implementation was first proposed in [84] for saving on computational complexity. Differences between time-varying and time-invariant Kalman filter implementations in the context of speech enhancement are studied in [94]. The difference between the time-variant Kalman filter (5) and the time-invariant implementation (23-25) is attributed to the fact that the former approach enables accurate modelling of the transients at frame boundaries. In Figure 14, the time-invariant Kalman gain is plotted against the time-variant gain for a voiced speech segment. The first element of the Kalman gain vectors is used in the plot. When k is small, the time-varying Kalman gain is "large" in order to obtain a fast decay of the transient, whereas the gain decreases with time so that the variance is small as well.

Next, we discuss the difference between the Kalman smoother and the Kalman filter. Since the noisy measurement set available to the filter, is a subset of measurements, available to the smoother, the obvious relation holds:

$$E\{(\mathbf{s}_k - \hat{\mathbf{s}}_k^S)(\mathbf{s}_k - \hat{\mathbf{s}}_k^S)^T\} \leq E\{(\mathbf{s}_k - \hat{\mathbf{s}}_k^F)(\mathbf{s}_k - \hat{\mathbf{s}}_k^F)^T\} \qquad (26)$$

However, this relation does not tell much of the perceptual differences between the two algorithms, which is of greater importance. This topic is investigated in paper B of this thesis.

The efficient implementation of the Kalman-fixed interval smoother is based on the Rauch-Tung-Stribel recursion [95]. Let the index 0 is assigned to the current speech sample, and the smoother delay is M samples. This leads to a "two-pass" algorithm. First, we run the Kalman filter over the interval $[0, M]$ and for each time instant k collect the values $\hat{\mathbf{x}}_{k|k-1}$, $\mathbf{P}_{k|k-1}$ and $\mathbf{P}_{k|k}$. Then, the smoothed state estimates and the corresponding sample estimate are obtained in reversed order $k = M, M-1, \ldots, 0$, through the recursion:

$$\begin{aligned} \hat{\mathbf{x}}_{k-1|M} &= \hat{\mathbf{x}}_{k-1|k-1} + \mathbf{P}_{k-1|k-1}\mathbf{F}^T\mathbf{P}_{k|k-1}^{-1}[\hat{\mathbf{x}}_{k|M} - \hat{\mathbf{x}}_{k|k-1}] \\ \hat{s}_{k-1} &= \mathbf{H}^T\hat{\mathbf{x}}_{k-1|M}. \end{aligned}$$

In paper B, we use Bryson-Frazier recursion [79], which is alternative to the outlined Rauch-Tung-Stribel algorithm. The Bryson-Frazier recursion is selected in the paper to facilitate the presentation and has no practical advantages over the Rauch-Tung-Stribel recursion.

3.2 Perceptually Motivated Algorithms

In the last decade, researchers have turned their attention to integrating the available knowledge of the human auditory system into noise suppression algorithms.

The study presented in [96] is focused on attenuation of musical noise, produced by the signal subspace speech enhancement algorithms. The basic concept is to place a perceptual-postfilter at the output of the signal subspace algorithm. This postfilter utilizes properties of the human auditory system, in an attempt to attenuate the residual noise with minimal speech distortion. The residual noise attenuation is based on an estimate of the masking threshold function.

Approaches to incorporate properties of the human auditory system directly into signal subspace methods or in subtraction based methods are presented in [97] and [98] respectively. A similar formulation the perceptual postfilter is used to further enhance the output of a Kalman filter-based noise suppression system [99], [100].

An estimate of the masking threshold function is used to control the parameters of a subtractive type noise suppression system in [101] and [102].

The perceptually motivated approach for speech enhancement, proposed in [103], avoids calculation of the masking threshold function. Instead, the method integrates the perceptual weighting technique, used in CELP coding [104], with subtractive type noise suppression algorithm.

4 Post-Processing Techniques Speech Enhancement Techniques

In addition to the discussed pre-processor speech enhancement techniques that aim at attenuating acoustic background noise, speech enhancement can be achieved by a post-processor, Fig. 1. Typically, the purpose of the post-processor speech enhancement processing is to attenuate the quantization noise in the synthesized speech signal.

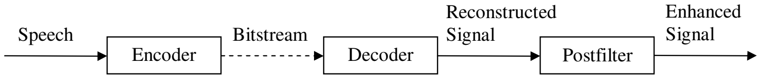

Figure 15: Configuration speech codec - speech enhancement post-processor.

In a speech decoder the synthesized speech is typically processed by a formant postfilter that emphasizes the formant frequencies and deemphasizes the valleys in between [105]. Additionally the synthesized speech can be processed by a pitch postfilter [106]. The purpose of a pitch postfilter is to emphasize frequency components at pitch harmonic peaks.

4.1 Theoretical Motivation

The existence of a postfilter at the speech decoder can be motivated formally by rate-distortion theory. This theory indicates that encoding at low bit rates with respect to squared-error distortion will result in a decoded signal with a spectrum different from that of the original signal [107], [108]. This theoretical result is often referred to as reverse water-filling. It suggests that the synthesis filter should differ from signal model filter. The presented in this section relations are valid under Gaussian assumption, but that we assume the basic principles carry over to speech signals.

The operation of the postfilter can be understood from graphs in power-spectral domain. Let λ be an auxiliary variable the control the operating point of an ideal coder. The area below both λ and the power spectrum $P(\omega)$ defines the distortion, see Fig. 16. Since the reconstructed signal and the quantization error are independent in an ideal codec, the sum of the distortion and the power spectrum of the reconstructed signal forms the power spectrum of the original signal. The relationship between the power

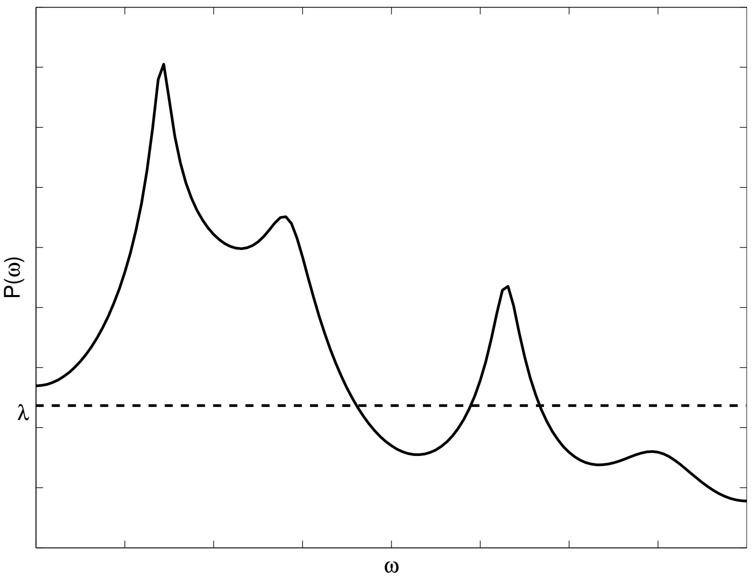

Figure 16: The reverse water-filling principle.

spectra of the reconstructed signal $\hat{s}$ and original signal s is

$$P_{\hat{s}}(\omega) = \max\left(P_s(\omega) - D(R), 0\right), \tag{27}$$

where distortion is denoted by D, and rate by R

$$\begin{aligned} R &= \frac{1}{4\pi}\int_{-\pi}^{+\pi} \max\left[0, \log\left(\frac{P_s(\omega)}{\lambda}\right)\right] d\omega & (28) \\ D &= \frac{1}{2\pi}\int_{-\pi}^{+\pi} \min\left[\lambda, P_s(\omega)\right] d\omega, & (29) \end{aligned}$$

Despite of the fact that postfilters are historically designed to reduce the perceived loudness of the excess noise in spectral valleys, in the light of reverse water-filling theory, the postfilters can be considered as an approximate implementation of the difference between a signal model filter and a synthesis filter.

4.2 Long- and Short-Term Postfiltering

There are two main types of postfilters. A formant postfilter reduces the effect of quantization noise by emphasizing the formant frequencies and deemphasizing the spectral valleys, while a pitch postfilter aims at emphasizing frequency components at pitch harmonic peaks.

The motivation for the postfiltering function arises from knowledge of the human auditory system, and particularly the concept of signal masking. In general, the masking threshold has a peak at the frequency of the tone, and monotonically decreases on both sides of the peak. This means that the noise components near the tone frequency (speech formants) are allowed to have higher intensities than other noise components that are farther away (spectrum valleys).

Psychoacoustical experiments show that the speech formants are much more important than spectral valleys, and the intensity of the spectral valleys can be significantly attenuated, without causing an audible distortion [109]. Therefore, by attenuating the signal component in spectral valleys, the postfilter only introduces minimal perceived distortion in the speech signal, still achieving noise reduction.

A general formant postfilter is given by a pole-zero filter [106]:

$$H_s(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}. \tag{30}$$

$A(z/\gamma) = 1 + \sum_{k=1}^{p} a_k (z/\gamma)^{-k}$ is the adaptive short term prediction-error filter, γ_1 and γ_2 are fixed parameters that control the degree of spectral emphasis, $0 < \gamma_1 < \gamma_2 < 1$, and p is the order of LP analysis, typically set to ten.

A problem with the basic formant postfilter of equation (30) is that it generally has a low-pass character, and the processed speech sounds muffled. It is desirable to develop a formant postfilter that has no spectral tilt. $H_t(z)$ is a tilt correction filter of the form [106]:

$$H_t(z) = (1 - \mu\ z^{-1}), \tag{31}$$

and it is controlled by the parameter μ that can be a function of the first reflection coefficient [110].

The energy of the synthesized signal is typically lower than the energy of the postfiltered signal. An adaptive gain control factor G_s compensates for the time-varying gain difference between the synthesized speech vector $\hat{\mathbf{s}}$ and the postfiltered speech vector $\hat{\mathbf{s}}_f$,

$$G_s = \sqrt{\frac{\hat{\mathbf{s}}^T \hat{\mathbf{s}}}{\hat{\mathbf{s}}_f^T \hat{\mathbf{s}}_f}}. \tag{32}$$

The gain is usually computed over 5 ms blocks, and linearly interpolated over time. Finally, the combined short-term postfilter can be expressed as:

$$H(z) = G_s H_s(z) H_t(z). \tag{33}$$

The postfilter parameters are set to different values values, dependent on the particular speech codec. For example in G.723.1 [111] $\gamma_1 = 0.65$, $\gamma_2 = 0.75$, and μ is a function of the firs reflection coefficient.

The most popular form of the pitch postfilter is described in [106]:

$$H_l(z) = G_l \frac{1 + \rho_1 z^{-\Lambda}}{1 - \rho_2 z^{-\Lambda}}, \tag{34}$$

where Λ is the pitch lag, the coefficients ρ_1 and ρ_2 control the gain of the pitch postfilter, and the overall gain G_l equalizes the energy of the input and output signals and is calculated similarly to the automatic gain control G_s in the formant postfilter.

The described formant and pitch postfiltering structure is not unique. A variant of code-excited linear prediction postfilter design technique that uses a frequency-domain approach, has been developed for sinusoidal coding systems [112]. This postfilter is a normalized, compressed version of the spectrally flattened vocal tract envelope. Let us define $R(\omega)$ by

$$\log R(\omega) = \log A(\omega) - \log T(\omega), \tag{35}$$

where $A(\omega)$ is the spectral envelope, and $T(\omega)$ is a first-order all-pole model of the spectrum tilt:

$$T(\omega) = \frac{1}{1 - a_1 e^{-j\omega}}. \tag{36}$$

a_1 is defined as the coefficient in the first order LP analysis, i.e., ratio between the first and zeroth order correlation coefficients. Then $R(\omega)$ is normalized to have unit gain, and root-γ compression rule is applied, with $\gamma \in (0, 1)$

$$\tilde{R}(\omega) = \left[\frac{R(\omega)}{R_{\max}}\right]^{\gamma} \tag{37}$$

Both the formant and pitch postfilters are still open research topics. Recent studies on the pitch postfilter can be found in [113], [114], and a novel form of the formant postfilter has been proposed recently in [115].

In [106] it was noted that in addition to quality enhancement of coded speech, the postfilter can be used for general speech enhancement. Experiments with the postfilter, or similar structures, in a general speech enhancement application can be found in [116–121]. In paper C we extend this idea by adapting the postfilter parameters to changing environment conditions. The adaptation is based on the advanced psychoacoustically motivated measure [3].

5 Summary of Contributions

The focus of this thesis is on quality assessment and enhancement of speech communication systems. The main contributions of the thesis can be summarized as follows: 1) explaining and solving the conflict between mean square error causal linear filters and human perception, 2) improving the postfiltering scheme used in speech coding, based on a psychoacoustically motivated distortion measure, and 3) proposing a novel, low-complexity, concept for non-intrusive speech quality assessment. Short summaries of the three papers included in the thesis are presented below.

Paper A: On Causal Algorithms for Speech Enhancement

Kalman filtering is a powerful technique for the estimation of a signal observed in noise that can be used to enhance speech observed in the presence of acoustic background noise. In a speech communication system, the speech signal is typically buffered for a period of 10 to 40 ms and, therefore, the use of either a causal or a noncausal filter is possible. We show that the causal Kalman algorithm is in conflict with the basic properties of human perception and address the problem of improving its perceptual quality. We discuss two approaches to improve perceptual performance. The first is based on a new method that combines the causal Kalman algorithm with pre- and postfiltering to introduce perceptual shaping of the residual noise. The second is based on the conventional Kalman smoother. We show that a short lag removes the conflict resulting from the causality constraint and we quantify the minimum lag required for this purpose. The results of our objective and subjective evaluations confirm that both approaches significantly outperform the conventional causal implementation. Of the two approaches, the Kalman smoother performs better if the signal statistics are precisely known, if this is not the case the perceptually weighted Kalman filter performs better.

Paper B: Low Complexity, Non-Intrusive Speech Quality Assessment

Monitoring of speech quality in emerging heterogeneous networks is of great interest to network operators. The most efficient way to satisfy such a need is through non-intrusive, objective speech quality assessment. In this paper we describe an algorithm for monitoring the speech quality over a network with extremely low complexity and memory requirements. The features used in the proposed algorithm can be computed from commonly used speech-coding parameters. Reconstruction and perceptual transformation of the signal is not performed. The critical advantage of the approach lies in generating quality assessment ratings without explicit distortion modelling. The results from the performed simulations indicate that the proposed output-based objective quality measure performs better than the ITU-T P.563 standard.

Paper C: Generalized Postfilter for Speech Quality Enhancement

Postfilters are commonly used in speech coding for attenuation of quantization noise. In the presence of acoustic background noise or distortion due to tandeming operations, the postfilter parameters are not adjusted and the performance is not optimal. We propose a modification that consists of replacing the non-adaptive postfilter parameters with parameters that adapt to variations in selected parameters obtained from the noisy speech, e.g., the spectral flatness. This generalization of the postfiltering concept can handle a larger number of distortions, but has the same computational complexity and memory requirements as the conventional postfilter. Test results indicate that the presented algorithms improve on the standard postfilter, as well as on the combination of a noise attenuation pre-processor and the conventional postfilter.

References

[1] B. C. J. Moore, *An Introduction to the Psychology of Hearing.* London: Academic Press, 1989.

[2] E. Zwicker and H. Fastl, *Psycho-acoustics: Facts and Models.* New York: Springer Verlag, 1999.

[3] T. Dau, D. Püschel, and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system. I. Model structure,” *J. Acoust. Soc. Am.*, vol. 99, pp. 3615–3622, 1996.

[4] A. Rix, A. Bourret, and M. Hollier, “Models of human perception,” *BT Technology Journal*, vol. 17, pp. 24–34, 1999.

[5] S. Voran, “A simplified version of the ITU algorithm for objective measurement of speech codec quality,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 1998, pp. 537–540.

[6] H. Knagenhjelm and W. B. Kleijn, “Spectral dynamics is more important than spectral distortion,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 1995, pp. 732–735.

[7] F. Norden and T. Eriksson, “Time evolution in LPC spectrum coding,” *IEEE Trans. Speech, Audio Processing*, vol. 12, pp. 290–301, 2004.

[8] T. Quatieri and R. Dunn, “Speech enhancement based on auditory spectral change,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 2002, pp. 257–260.

[9] G. Schuller, B. Yu, D. Huang, and B. Edler, “Perceptual audio coding using adaptive pre- and post-filters and lossless compression,” *IEEE Trans. Speech, Audio Processing*, vol. 10, pp. 379–390, 2002.

[10] M. Schroeder, B. Atal, and J. Hall, “Optimizing digital speech coders by exploiting the masking properties of the human ear,” *J. Acoust. Soc. Am.*, vol. 66, pp. 1647–1652, 1979.

[11] W. Chang and C. Wang, "Audio coding using masking-threshold adapted perceptual filter," in *Proc. IEEE Workshop Speech Coding for Telecom.*, 1993, pp. 13–15.

[12] G. Charestan, R. Heusdens, and S. van der Par, "A gammatone-based psychoacoustical modeling approach for speech and audio coding," in *Proc. IEEE Workshop Circuits, Systems, Signal Processing*, 2001, pp. 321–326.

[13] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Selected Areas in Commun.*, vol. 6, pp. 314–323, 1988.

[14] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis.* Amsterdam: Elsevier Science Publishers, 1995.

[15] T. Painer and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, pp. 451–513, 2000.

[16] J. Plasberg and W. B. Kleijn, "The Sensitivity Matrix: Using Advanced Auditory Models in Speech and Audio Processing," *IEEE Trans. Speech, Audio Processing*, to appear.

[17] P. Hedelin, F. Norden, and J. Skoglund, "SD optimization of spectral coders," in *Proc. IEEE Workshop on Speech Coding*, 1999, pp. 28–30.

[18] R. Reynolds and A. Rix, "Quality VoIP - an engineering challenge," *BT Technology Journal*, vol. 19, pp. 23–32, 2001.

[19] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," 1996.

[20] ITU-T Rec. P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," 1996.

[21] ITU-R Rec. BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems," 2001.

[22] ITU-R Rec. BS.562-3, "Subjective assessment of sound quality," 1990.

[23] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 2, 1977, pp. 204–207.

[24] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Communication*, vol. 16, pp. 225–244, 1995.

[25] ITU-T. Rec. P.810, "Modulated Noise Reference Unit," 1996.

[26] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," International Telecommunication Union, 1998.

[27] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality.* Prentice Hall, 1988.

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

[29] J. Beerends and J. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc*, vol. 42, no. 3, pp. 115–123, 1994.

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

[31] ITU-R Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2005.

[32] ITU-R. BS.1387, "Method for Objective Measurements of Perceived Audio Quality (PEAQ)," 1998.

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

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

[35] W. Zha and W.-Y. Chan, "Objective speech quality measurement using statistical data minimg," *Journal Applied Signal Process.*, vol. 9, pp. 1410–1424, 2005.

[36] D. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech, Audio Processing*, vol. 13, pp. 821–831, 2005.

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

[38] G. Chen and V. Parsa, "Bayesian model based non-intrusive speech quality evaluation," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 2005, pp. 385–388.

[39] T. Falk, Q. Xu, and W.-Y. Chan, "Non-intrusive GMM-based speech quality measurement," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 2005, pp. 125–128.

[40] D. Picovici and A. Mahdi, "Output-based objective speech quality measure using self-organizing map," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 2003, pp. 476–479.

[41] J. Liang and R. Kubichek, "Output-based objective speech quality," *IEEE 44th Vehicular Technology Conf.*, vol. 3, no. 8-10, pp. 1719–1723, 1994.

[42] ITU-T P. 563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," International Telecommunication Union, Geneva, Switzerland, 2004.

[43] O. Au and K. Lam, "A novel output-based objective speech quality measure for wireless communication," *Signal Processing Proceedings, 4th Int. Conf.*, vol. 1, pp. 666–669, 1998.

[44] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process*, vol. 13, pp. 600–612, 2004.

[45] ITU-T Rec. G.107, "The e-model, a computational model for use in transmission planning," 2003.

[46] M. Werner, T. Junge, and P. Vary, "Quality control for AMR speech channels in GSM networks," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 3, 2004, pp. 1076–1079.

[47] J. Lim, Ed., *Speech Enhancement.* Englewood Cliffs, NJ: Prentice Hall, 1983.

[48] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Processing*, vol. 9, pp. 504–512, 2001.

[49] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 3, 2000, pp. 1875–1878.

[50] R. Bouquin-Jeannes and G. Faucon, "Proposal of a voice activity detector for noise reduction," *Electronics Lett.*, vol. 30, pp. 930–932, 1994.

[51] Y. Cho, K. Al-Naimi, and A. Kondoz, "Imrpved voice activity detection based on a smoothed statistical likelihood ratio," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 2, 2001, pp. 737–740.

[52] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 1998, pp. 365–368.

[53] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 113–120, 1979.

[54] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 345–349, 1994.

[55] Z. Goh, K. Tan, and B. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech, Audio Processing*, vol. 6, pp. 287–292, 1998.

[56] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtraction (NSS), hidden markov models and the projection for robust speech recognition in cars," *Speech Commun.*, vol. 11, pp. 215–228, 1992.

[57] P. Handel, "Low-distortion spectral subtraction for speech enhancement," in *Proc. Eurospeech*, 1995, pp. 1549–1552.

[58] Y. Ephraim and H. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 251–266, 1995.

[59] R. Vetter, "Single channel speech enhancement using MDL-based subspace approach in bark domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2001, pp. 641–644.

[60] P. Bodin and L. Villemoes, "Spectral subtraction in the time-frequency domain using waveletpackets," in *Proc. IEEE Workshop Speech Coding for Telecom.*, 1997, pp. 47–48.

[61] S. Chang, Y. Kwon, S. Yang, and I. Kim, "Speech enhancement for nonstationary noise environment by adaptive wavelet packet," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 2002, pp. 561–564.

[62] N. Whitmal, J. Rutledge, and J. Cohen, “Wavelet-based noise reduction,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 5, 1995, pp. 3003–3006.

[63] D. Anderson and M. Clements, “Audio signal noise reduction using multi-resolution sinusoidal modeling,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 2, 1999, pp. 805–808.

[64] J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, pp. 197–210, 1978.

[65] T. Sreenivas and P. Kirnapure, “Codebook constrained Wiener filtering for speech enhancement,” *IEEE Trans. Speech, Audio Processing*, vol. 4, no. 5, pp. 383–389, 1996.

[66] J. Hansen and M. Clements, “Constrained iterative speech enhancement with application to speech recognition,” *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, 1991.

[67] Y. Ephraim, “A Bayesian estimation approach for speech enhancement using hidden Markov models,” *IEEE Trans. Signal Processing*, vol. 40, pp. 725–735, 1992.

[68] ——, “A minimum mean square error approach for speech enhancement,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 2, 1990, pp. 829–832.

[69] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. Speech Audio Processing*, 2004, accepted for publication.

[70] ——, “Speech enhancement using a-priori information,” in *Proc. Eurospeech*, vol. 2, 2003, pp. 1405–1408.

[71] TIA/EIA/IS-127, “Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems,” 1997.

[72] 3GPP2 C.S0030-0, “Selectable Mode Vocoder Service Option for Wideband Spread Spectrum Communication Systems,” 2001.

[73] T. Ohya, H. Suda, and T. Miki, “5.6 kbits/s PSI-CELP of the half-rate PDC speech coding standard,” in *Proc. IEEE 44th Vehicular Technology Conference*, vol. 3, 1994, pp. 1680–1684.

[74] 3GPP TR 26.978, “Results of the Adaptive Multi-Rate (AMR) Noise Suppression Selection Phase,” 2001.

[75] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 443–445, 1985.

[76] R. McAulay and M. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 137–145, 1980.

[77] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109–1121, 1984.

[78] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 3, 2000, pp. 1479–1482.

[79] T. Kailath, A. Sayed, and B. Hassiby, *Linear Estimation.* New Jersey: Prentice Hall, 2000.

[80] T. Soderstrom, *Discrete-time Stochastic Systems*, 2nd ed. London: Springer-Verlag, 2002.

[81] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* New York: Wiley, 1949.

[82] T. Kailath, *Lectures on Wiener and Kalman Filtering.* New York: Springer Verlag, 1981.

[83] M. Hayes, *Statistical Digital Signal Processing and Modeling.* US: Wiley, 1996.

[84] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 12, 1987, pp. 177–180.

[85] J. Gibson, B. Koo, and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Processing*, vol. 39, pp. 1732–1742, 1991.

[86] M. Gabrea, "An adaptive Kalman filter for the enhancement of speech signals," in *Proc. Int. Conf. Spoken Language Processing*, 2004, pp. 2709–2713.

[87] P. Sorqvist, P. Handel, and B. Ottersten, "Kalman filtering for low distortion speech enhancement in mobile communication," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 2, 1997, pp. 1219–1222.

[88] V. Grancharov, S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Robust spectrum quantization for LP parameter enhancement," in *Proc. XII European Signal Processing Conf.*, 2004, pp. 1951–1954.

[89] W. Wen-Rong and C. Po-Cheng, "Subband Kalman filtering for speech enhancement," *IEEE Trans. Circuits and Systems II: Analog, Digital Signal Processing*, vol. 45, pp. 1072–1083, 1998.

[90] K. Lee and S. Jung, "Time-domain approach using multiple kalman filters and EM algorithm to speech enhancement with nonstationary noise," *IEEE Trans. Speech, Audio Processing*, vol. 8, pp. 282–291, 2000.

[91] S. Gannot, D. Burshtein, and E.Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 4, pp. 373–385, 1998.

[92] Z. Goh, K.-C. Tan, and B. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. Speech, Audio Processing*, vol. 7, no. 5, pp. 510–524, 1999.

[93] M. Niedzwiecki and K. Cisowski, "Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals," *IEEE Trans. Signal Processing*, vol. 44, pp. 528–537, 1996.

[94] D. Popescu and I. Zeljkovic, “Kalman filtering of colored noise for speech enhancement,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 12, 1998, pp. 997–1000.

[95] B. D. O. Anderson and J. B. Moore, *Optimal Filtering.* Englewood Cliffs, NJ: Prentice Hall, 1979.

[96] M. Klein and P. Kabal, “Signal subspace speech enhancement with perceptual post-filtering,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, Orlando, USA, 2002, pp. 537–540.

[97] F. Jabloun and B. Champagne, “Incorporating the human hearing properties in the signal subspace approach for speech enhancement,” *IEEE Trans. Speech, Audio Processing*, vol. 11, pp. 700–708, 2003.

[98] S. Gustafsson, P. Jax, and P. Vary, “A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 1998, pp. 397–400.

[99] N. Ma, M. Bouchard, and R. Goubran, “Perceptual Kalman filtering for speech enhancement in colored noise,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, Montreal, Canada, 2004, pp. 17–21.

[100] ——, “A perceptual Kalman filtering-based approach for speecn enhancement,” in *Proc. Seventh International Symposium on Signal Processing and its Applications*, vol. 1, 2003, pp. 373–376.

[101] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech, Audio Processing*, vol. 7, pp. 126–137, 1999.

[102] D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, “Speech enhancement based on audible noise suppression,” *IEEE Trans. Speech, Audio Processing*, vol. 5, pp. 497–514, 1997.

[103] Y. Hu and P. Loizou, “A perceptually motivated approach for speech enhancement,” *IEEE Trans. Speech, Audio Processing*, vol. 11, pp. 457–465, 2003.

[104] M. Schroeder and B. Atal, “Code-excited linear prediction (CELP): high-quality speech at very low bit rates,” in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 10, 1985, pp. 937–940.

[105] V. Ramamoorthy, N. Jayant, R. Cox, and M. Sondhi, “Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback,” *IEEE J. Select. Areas Commun.*, vol. 6, pp. 364–382, 1988.

[106] J.-H. Chen and A. Gersho, “Adaptive postfiltering for quality enhancement of coded speech,” *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 59–71, 1995.

[107] S. Andersen and W. B. Kleijn, “Reverse water-filling in predictive encoding of speech,” in *Proc. IEEE Workshop on Speech Coding*, vol. 3, 1999, pp. 105–107.

[108] W. B. Kleijn, *A Basis for Source Coding.* not published, 2006.

[109] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 115–131, 1994.

[110] W. B. Kleijn, D. Krasinski, and R. Ketchum, "Improved speech quality and efficient vector quantization in SELP," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, 1988, pp. 155–158.

[111] ITU-T. Rec. G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s," 1996.

[112] R. McAulay, T. Parks, T. Quatieri, and M. Sabin, *Sine-wave amplitude coding at low data rates.* in Advances in Speech Coding, B. Atal and V. Cuperman, and A. Gersho, Eds., Kluwer Academic Publishers, 1991.

[113] W. B. Kleijn, "Enhancement of coded speech by constrained optimization," in *Proc. IEEE Workshop on Speech Coding*, 2002, pp. 163–165.

[114] 3GPP TS 26.290, "Extended Adaptive Multi-Rate - Wideband (AMR-WB+) Codec; Transcoding Functions."

[115] W.-Y. Chen, P. Kabal, and T. Shabestary, "Perceptual postfilter estimation for low bit rate speech coders using Gaussian mixture models," in *Proc. Interspeech*, 2005, pp. 3161–3164.

[116] M. Zilovic, R. Ramachandran, and R. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer function," *IEEE Trans. Speech, Audio Processing*, vol. 6, pp. 260–267, 1998.

[117] R. Chandran and D. Marchok, "Compressed domain noise reduction and echo suppression for network speech enhancement," in *Circuits and Systems, Proc. of the 43nd Midwest Symp.*, vol. 1, 2000, pp. 10–13.

[118] H. Taddei, C. Beaugeant, and M. de Meuleneire, "Noise reduction on speech codec parameters," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 2004, pp. 497–500.

[119] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Noise-dependent postfiltering," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, 2004, pp. 457–460.

[120] P. Kabal, F. Wang, D. O'Shaughnessy, and R. Ramachandran, "Adaptive postfiltering for enhancement of noisy speech in the frequency domain," in *IEEE Int. Symp. Circuits and Systems*, 1991, pp. 312–315.

[121] R. Conway, T. Sreenivas, and R. Niederjohn, "Adaptive postfiltering applied to speech in noise," in *Circuits and Systems, 1989., Proc. of the 32nd Midwest Symp.*, 1989, pp. 101–104.