Thesis for the degree of Doctor of Philosophy

Improving Quality of Service in Baseband Speech Communication

Petko Nikolov Petkov

KTH Electrical Engineering

Communication Theory
School of Electrical Engineering
KTH – Royal Institute of Technology

Stockholm 2014
Petkov, Petko Nikolov
Improving Quality of Service in Baseband Speech Communication

Copyright ©2014 Petko Nikolov Petkov except where otherwise stated. All rights reserved.


Communication Theory
School of Electrical Engineering
KTH – Royal Institute of Technology
SE-100 44 Stockholm, Sweden
Abstract

Speech is the most important communication modality for human interaction. Automatic speech recognition and speech synthesis have extended further the relevance of speech to man-machine interaction. Environment noise and various distortions, such as reverberation and speech processing artifacts, reduce the mutual information between the message modulated in the clean speech and the message decoded from the observed signal. This degrades intelligibility and perceived quality, which are the two attributes associated with quality of service. An estimate of the state of these attributes provides important diagnostic information about the communication equipment and the environment. When the adverse effects occur at the presentation side, an objective measure of intelligibility facilitates speech signal modification for improved communication.

The contributions of this thesis come from non-intrusive quality assessment and intelligibility-enhancing modification of speech. On the part of quality, the focus is on predictor design for limited training data. Paper A proposes a quality assessment model for bounded-support ratings that learns efficiently from a limited amount of training data, scales easily with the sampling frequency, and provides a platform for modeling variations in the individual subjective ratings. The predictive performance of the model for the mean of the subjective quality ratings compares favorably to the state-of-art in the field. Patterns in the spread of the individual ratings are captured in the feature space of the training data.

Paper B focuses on enhancing predictive performance for the mean of the quality variable when the signal feature space is sparsely sampled by the training data. Using a Gaussian Processes framework, the deterministic signal-based feature set is augmented with a stochastic feature that is hypothesized to be jointly distributed with the target quality rating. An uncertainty propagation mechanism ensures that the variance of this feature is reflected in the prediction. The proposed architecture can take advantage of i) data that cannot be pooled due to subjective test protocol incompatibility and ii) models trained on data that are no longer available.

With respect to intelligibility enhancement, a hierarchical perspective of the speech communication process, extended from foundational work in the field, is used in paper C to create a unified framework for method analysis.
and comparison. A high-level intelligibility measure related to the probability for correct recognition is derived using a hit-or-miss distortion criterion in the transcription domain. The measure is used to optimize two speech modifications at different levels of the message encoding hierarchy leading to significantly enhanced intelligibility in noise. The conceptual novelty of the method comes at the cost of higher complexity and the requirement for additional information including message transcription, sound segmentation, and a model of speech.

Mapping the high-level measure to a lower level takes away the need for additional information and preserves asymptotically high-level optimality. Two methods are proposed to reduce degradation in the accuracy of the spectral dynamics due to additive noise. The focus of paper D is dynamics preservation in a range that is lower-bounded by an optimal band-power threshold. The performance of the method is competitive but allows for improvement in power efficiency. This issue is addressed in paper E which proposes and optimizes a distortion measure for spectral dynamics leading to a significant increase in intelligibility. Use of functional optimization techniques allows for families of solutions, among which are dynamic range compressors adaptive to the statistics of the speech and the noise.

**Keywords**: non-intrusive quality assessment, intelligibility enhancement, speech adaptation, speech modification, spectral dynamics distortion, automatic speech recognition, Gaussian Processes, dynamic range compression.
List of Papers

The thesis is based on the following papers:


In addition to papers A–E, the following papers and patent application have been produced in part by the author of the thesis:


Acknowledgements

I am grateful to my academic advisor Prof. W. Bastiaan Kleijn for giving me this opportunity to learn and create. When writing this thesis I could recall the individual discussions, physical or not, that have shaped this work.

I thank Jan Plasberg, my MS thesis supervisor, for waking up the interest in the field, Prof. Arne Leijon for his support and thought-provoking comments, Prof. Markus Flierl for the discussions we have had and Prof. Mikael Skoglund for the warm welcome to the Communication Theory lab.

I appreciate the efforts of my collaborators that have resulted in the improved quality of this work. In particular I would like to thank Dr. Jesus Vicente de Pena, Dr. Hannes Helgasson, Dr. Gustav Eje Henter and Dr. Mengqiu Zhang.

I am grateful to my colleagues from the Sound and Image Processing and the Communication Theory labs for creating a stimulating working and social environment that could sustain discussions on any topic from horlogerie haute and ice fishing to philosophy and, occasionally, work.

A number of friends have made these years even more enjoyable. Among them Agata and Konrad Hofbauer, Miroslav Petrov, Volodya Grancharov, Polina and Valyo Georgievi, Nina Burge, Olga Pleijte, Asa Danielsson, Grazyna Molin, Maria Jose Garcia, Gustavo Salcedo, Dr. Giorgio Regnoli, Pino Dangiola and the innebandy team that doesn’t stop to amaze me with their energy.

I am indebted to my family for their support and understanding.

Anick, thank you for your love and patience!

Petko Nikolov Petkov
Stockholm, May 2014
Contents

Abstract
List of Papers
Acknowledgements
Contents
Acronyms

I Introduction

1 QoS in Speech Communication

1.1 Subjective Evaluation of Quality and Intelligibility

1.1.1 Evaluation of Quality

1.1.2 Evaluation of Intelligibility

1.2 Models for Predicting Quality

1.2.1 Intrusive QA

1.2.2 Non-Intrusive QA

1.2.3 Parametric QA

1.3 Models for Predicting Intelligibility

1.3.1 Low-level methods

1.3.2 High-level methods

1.4 Intelligibility Enhancing Speech Modification

1.4.1 Speech Modification Strategies

2 Thesis Directions

2.1 Quality Assessment for Limited Data

2.1.1 Statistical Models for a bounded-support random variable

2.1.2 Prediction Augmentation with Correlated Data

2.2 Intelligibility Enhancing Speech Modification
3 System operation ........................................ C10
  3.1 Principle of Operation ............................... C10
  3.2 Signal Modifications ............................... C12
  3.3 Optimization Complexity and Memory Requirements C13
  3.4 System Implementation Considerations .......... C14
4 Experimental Results ................................. C15
  4.1 ASR Training ....................................... C15
  4.2 Reference System ................................. C16
  4.3 Listening Test Set-up .............................. C16
  4.4 Subjective Evaluation Results .................... C17
  4.5 Method Behavior .................................. C18
5 Conclusions ............................................ C21
References ............................................... C24

D Preservation of Speech Spectral Dynamics Enhances Intel-
ligibility D1
1 Introduction .......................................... D1
2 Theoretical foundations .............................. D3
  2.1 Conditions for preserving spectral dynamics ...... D3
  2.2 Modification parameter optimization for SDP ...... D5
3 Implementation ........................................ D6
4 Experimental Results ................................. D7
5 Conclusions ............................................ D10
References ............................................... D11

E Spectral Dynamics Recovery for Enhanced Speech Intelli-
gibility in Noise E1
1 Introduction .......................................... E1
2 Theoretical Foundations .............................. E4
  2.1 Single-band dynamics recovery .................... E5
    2.1.1 Problem formulation ............................ E5
    2.1.2 General solution to the single-band optim-
        ization problem ................................ E7
    2.1.3 Initial and final conditions .................... E8
    2.1.4 Parametrization of the final condition .... E8
  2.2 Multi-band dynamics recovery ................... E9
    2.2.1 Optimization problem ......................... E9
    2.2.2 Feed-through behavior in noise-free conditions E12
3 Practical Considerations ............................ E13
  3.1 Density estimation ................................ E13
  3.2 Solution biasing .................................. E14
  3.3 Setting the v parameter ............................ E15
  3.4 Algorithm .......................................... E15
4 Method Evaluation .................................... E16
## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACR</td>
<td>Absolute Category Rating</td>
</tr>
<tr>
<td>CCR</td>
<td>Comparison Category Ratings</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Density Function</td>
</tr>
<tr>
<td>CMOS</td>
<td>Comparison Mean Opinion Score</td>
</tr>
<tr>
<td>CSII</td>
<td>Coherence Speech Intelligibility Index</td>
</tr>
<tr>
<td>DCM</td>
<td>Discrete Choice Models</td>
</tr>
<tr>
<td>DCR</td>
<td>Degradation Category Ratings</td>
</tr>
<tr>
<td>DMOS</td>
<td>Degradation Mean Opinion Score</td>
</tr>
<tr>
<td>DRC</td>
<td>Dynamic Range Compression</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>GP</td>
<td>Gaussian Processes</td>
</tr>
<tr>
<td>HNM</td>
<td>Harmonic plus Noise Model</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electric and Electronics Engineers</td>
</tr>
<tr>
<td>IM</td>
<td>Intelligibility Model</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>ITFS</td>
<td>Ideal Time-Frequency Segregation</td>
</tr>
<tr>
<td>MARS</td>
<td>Multivariate Adaptive Regression Splines</td>
</tr>
<tr>
<td>MDL</td>
<td>Minimum Description Length</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Scores</td>
</tr>
</tbody>
</table>

 xi
MTF Modulation Transfer Function
MUSHRA Multiple Stimuli with Hidden Reference and Anchor
NB Narrow-Band
NCM Normalized Covariance Metric
NH Normal Hearing
NN Neural Network
ODE Ordinary Differential Equation
PDF Probability Density Function
PLP Perceptual Linear Prediction
PSD Power Spectral Density
QA Quality Assessment
QoS Quality of Service
RMS Root Mean Squared
SDR Signal to Distortion Ratio
SII Speech Intelligibility Index
SNR Signal to Noise Ratio
SS Spectral Shaping
STI Speech Transmission Index
STOI Short-Time Objective Intelligibility
SVR Support Vector Regression
TASL Transaction on Audio Speech and Language Processing
TF Time-Frequency
TTS Text-to-Speech
VAD Voice Activity Detection
VCV Vowel-Consonant-Vowel
WB Wide-Band
Part I

Introduction
Introduction

1 QoS in Speech Communication

Speech represents a communication modality where information exchange is achieved through the modulation of acoustic signals. With the advances in technology, the relevance of speech communication has extended beyond the scenario of human-to-human communication. Automatic speech recognition (ASR) and speech synthesis, or text-to-speech (TTS), have enabled speech communication between man and machine. Ideal conditions, however, are seldom the reality of a real-world communication scenario. Various environmental factors lower the capacity of the communication channel. In addition, degradation in perceived quality results in dissatisfaction and annoyance on the part of the human and reduces further the possibility for successful communication. This thesis is about speech, the evaluation of its quality and the enhancement of its intelligibility.

Quality of Service (QoS) is an indicator of the overall performance of a communication service and, applied to speech, reflects the state of two of its attributes - quality and intelligibility. Factors that degrade quality and intelligibility include, among others, environment noise [1], reverberation [2], non-linear artifacts from speech processing algorithms [3] and distortions caused by resource sharing in the communication equipment [4]. Continuously monitoring QoS provides key information regarding the characteristics of the communication channel and, with it, the means for adapting the operating parameters of the communication equipment with the objective of enhancing the end-user experience. Representative examples of adaptation include, e.g., the automated tuning of hearing aids [5], scaling in advanced speech and audio coders [6] and speech modification for enhanced intelligibility prior to presentation in noisy environments [7].

Since the objective of communication is transferring information, intelligibility can be viewed as the more fundamental of the two attributes. Thus, quality is primarily of interest when the intelligibility of the speech signal is sufficiently high to enable successful communication. Given the ability of normal-hearing subjects to understand speech even when the power of
the disturbance is significantly higher than that of the speech signal, the range of conditions where it is physically meaningful to evaluate quality is extensive.

Quality and intelligibility are quantified using numerical scales. In the case of quality, the scales are categorical and ordinal, and quantize the perceived deviation from an internal or a given (physical) reference. Intelligibility, on the other hand, is represented by a continuous number, lower-bounded by zero and upper-bounded by one, that can be thought of as the proportion of the correctly identified and the entire information.

The most reliable approach to measuring quality and intelligibility is through subjective experiments. The cost of organizing such experiments, however, is high and the process is time-consuming. As a result, there is a vivid interest in the development of predictive objective models for the two attributes. Experimental data from subjective listening experiments is currently indispensable from model selection and design. The gradual increase in knowledge of the functioning of the auditory system at the neural level [8, 9, 10, 11] is expected to reduce the dependence on the data, in particular, with regards to intelligibility. Significant individual preferences in the quality domain, e.g. [12], suggest that data dependence in quality prediction is difficult to alleviate. Even when the objective is predicting the average trends in perceived quality, model design without training data is not yet feasible. The efficient use of the training data in quality assessment is, thus, a major concern and is considered in [13, 14] (papers A and B).

A number of successful models for predicting intelligibility have been proposed in the literature [15, 16, 17]. The high correlation between the predictive scores of these models, for the intended operating conditions, and the corresponding subjective scores opens up the room for optimizing speech modifications with the objective of intelligibility enhancement. This is the procedure of preemptively modifying the clean speech signal before it is affected by intelligibility degrading adverse effects such that, after rendering, intelligibility is maximized. Adapting the modification to the type and the level of the disturbance, in theory, delivers maximum intelligibility gain with minimum processing artifacts and loss of naturalness. By far, the most common application scenario is that of additive noise. Intelligibility enhancing speech modification with application to additive noise is addressed in [18, 19, 20] (papers C, D and E).

The remainder of this section provides the background information necessary to prepare the reader for the presentation of the thesis research topics in Section 2. Subjective measurement of quality and intelligibility is reviewed in Section 1.1. The main advances in the development of models for quality and intelligibility prediction are presented in Section 1.2 and Section 1.3 respectively. Speech modification methods for intelligibility enhancement in a noisy environment are discussed in Section 1.4.
1.1 Subjective Evaluation of Quality and Intelligibility

Subjective data collected through listening tests is needed for the development and the validation of models for predicting quality and intelligibility. The listening tests are conducted according to standardized experimental protocols that address various evaluation scenarios. To provide an outlook on data collection for quality assessment (QA) and to illustrate how the proposed methods for intelligibility enhancement in noise are evaluated in practice, the main QA and intelligibility-related test protocols are discussed in Section 1.1.1 and Section 1.1.2 respectively.

1.1.1 Evaluation of Quality

In the context of speech processing, quality effectively characterizes the naturalness of speech. Deviations from what is perceived as natural degrade human perception of quality. External factors affecting the degree of deviation include, among others, additive noise, reverberation, echo, loss and delay of packets in packet switched networks. Internal factors are, e.g., inaccuracies and approximations in the production model of a text-to-speech (TTS) system [21].

Apart from the cause of degradation, another classification criterion is the communication scenario in which quality is measured. The three classes are: i) listening-only, ii) talking-only and iii) conversational. The focus here is on quality degradation due to external influences [22] in listening-only situations. This choice is motivated with the high practical importance of the considered scenario.

Obtaining consistent and comparable quality ratings requires the use of standardized test methods. A broad classification identifies absolute category ratings (ACR), and reference-based protocols such as degradation category ratings (DCR), comparison category ratings (CCR) [23] and multiple stimuli with hidden reference (MUSHRA) [24].

ACR testing employs the five-category rating scale shown in Table 1, where an ordinal sequence of numbers quantizes the perceived quality. More detailed scales are straightforward to implement, but the benefit from the higher accuracy is constrained by noise in the subjective ratings. When evaluating telecommunication equipment or the performance of a speech processing algorithm, it is typically the average trend over a large number of test participants that is of interest. As a result, the ACR mean, referred to

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Bad</td>
<td>Poor</td>
<td>Fair</td>
<td>Good</td>
<td>Excellent</td>
</tr>
</tbody>
</table>
as mean opinion score (MOS) [25], has become an established performance metric. Higher order moments characterizing the score distribution are generally neglected.

The DCR rating scale quantifies degradation with respect to a reference, which would typically be the clean signal. Similar to ACR, it uses a five-point scale as shown in Table 2. The CCR rating scale, shown in Table 3,

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annoying</td>
<td></td>
<td>Annoying</td>
<td>Annoying</td>
<td>Annoying</td>
<td>Inaudible</td>
</tr>
</tbody>
</table>

is two-sided and best-suited for the comparison of competing algorithms. Mean results are defined for both DCR (DMOS) and CCR (CMOS) [23],

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much</td>
<td>Slightly</td>
<td>About</td>
<td>Slightly</td>
<td>Much</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worse</td>
<td>Worse</td>
<td>Worse</td>
<td>the Same</td>
<td>Better</td>
<td>Better</td>
<td>Better</td>
<td></td>
</tr>
</tbody>
</table>

and used for evaluation in practice.

MUSHRA can be viewed as a hybrid between CCR and DCR as it includes both a reference and multiple processed versions of the signal. It is used primarily in audio quality assessment for characterizing the effect of small impairments. Given the higher complexity of this protocol, data collection is time-consuming and the results are harder to analyze.

The relative simplicity of the ACR rating scale facilitates data collection and explains the popularity of this protocol. An extensive amount of, primarily proprietary, see, e.g., [26, 27], but also some publicly available data [28, 29, 30] has been collected and used for the development and the validation of QA models.

1.1.2 Evaluation of Intelligibility

Intelligibility assessment builds on the evaluation of recognition rates of speech material in various test conditions. Different test protocols have been established depending on the purpose of the test. The characteristics commonly used to classify intelligibility tests are the adaptivity, the type and the extent of the test material [31].

Test adaptation refers to adjusting the presentation material based on
previous responses by the test subject. Progressively increasing the task difficulty by reducing the signal-to-noise ratio (SNR) has implications on efficiency when performing, e.g., assessment of hearing degradation. The choice of test material may target the evaluation of the intelligibility of particular sounds and sequences thereof [16, 32, 33] or overall intelligibility [34, 35] in terms of complete sentences. The extent of the test material identifies open-set tests using large vocabulary [35, 18] and closed-set tests using limited vocabulary [36, 37] which may or may not be presented to the participant before testing.

The choice of a particular test procedure depends on the objective. Evaluating the intelligibility of conversational speech rendered in noise with an open-set, non-adaptive test provides a good approximation to the expected performance in practice. Independent of the test type, the figure of merit commonly used to quantify intelligibility is the ratio of the correctly recognized and the total number of lexical units, e.g., [38], possibly excluding a particular subset due to high predictability [35].

1.2 Models for Predicting Quality

A comprehensive overview of models for quality assessment of speech and audio can be found in [39]. A more speech-centric perspective is provided in [22]. The objective of this section is to introduce the main concepts and facilitate the presentation of the related thesis contributions.

The need for quality assessment arises in different application scenarios and, as a result, three distinct types of model architectures are discussed in the literature. These are referred to as intrusive (reference-based) [40, 41, 42], non-intrusive [43, 26, 27] and parametric [44]. Intrusive and non-intrusive models operate by extracting features from the noisy and the clean or the noisy signals only and mapping a function of these to the domain of quality estimates. Parametric models, on the other hand, base the quality prediction on the values of indicators describing the operating conditions of the communication equipment. Thus, high delay and packet loss, for packet-based networks, and down-scaling the bit-rate of a speech coder translate into quality degradation. The mapping to quality predictions, for both signal-based and parametric models, is commonly performed by a regression [45, 46, 47], making such models well-suited for modeling quasi-continuous MOS random variable. Learning of the mapping function requires data from subjective experiments.

1.2.1 Intrusive QA

It is the general view that intrusive modeling of quality is a mature field at least in comparison to other types of QA model development [39, 22]. The principles of operation of some of the established models are reviewed in the
following, with the intent to provide a historical perspective and a platform for the discussion of the challenges encountered in non-intrusive methods.

Comparing features extracted from the clean and the corresponding noisy speech signals, intrusive QA models generate a numerical value that is expected to correlate well with subjective scores such as, e.g., MOS. The features can be derived from the waveform, spectral or auditory [48] representation of the signal. Auditory modeling takes into consideration psychoacoustic properties and offers the advantage of compact signal representations. As a result, the majority of the proposed QA models [49, 40, 41, 42] employ auditory features. Auditory information is not used, e.g., in [50], which is a model predicting the quality of reverberant speech.

The model stage performing feature comparison is called a cognitive model [51]. It computes a measure of similarity or distance between the noisy and the clean signal feature sets and, ideally, reflects the degradation of quality. The availability of the reference signal leads to a model architecture the performance of which is not heavily dependent on the amount of training data from subjective experiments.

From a signal processing perspective, a large proportion of QA models includes a frame-based signal analysis stage. Proper comparison between the distorted and the clean signals, especially in the context of delay and packet-loss impairments, requires adequate frame alignment and use of voice activity detection (VAD). More importantly, it is necessary to either i) derive (global) features representative of the entire utterance [43, 26] or ii) aggregate (local) frame-based quality predictions into a global quality prediction [27].

The output of a QA model does not necessarily belong to the range of the subjective quality ratings that it models. Another issue that needs to be taken into account when evaluating the performance of a QA model is the low consistency in the outcomes of subjective evaluation experiments. To alleviate the effect of these two factors, it is possible to employ an additional monotonic polynomial mapping of the raw quality predictions. This approach is applied to the model from [40], where the parameters of the polynomial mapping are optimized over a large amount of data [52]. Raw score manipulation can also be a part of the QA model architecture [42].

### 1.2.2 Non-Intrusive QA

The requirement for a clean reference signal limits the applicability of intrusive models to pre-deployment performance evaluation and off-line network diagnostics. Non-intrusive models are designed to predict quality without the need for a reference signal. Provided that it is no longer possible to perform accurate cognitive modeling, the task of predicting quality becomes more challenging. A review of existing models reveals that these, generally, have a larger number of free-parameters, compared to intrusive models, and
exhibit heavier dependence on the training data.

The model architectures encountered in the literature perform signal preprocessing, feature extraction (typically done on a frame-by-frame basis) and mapping from the features to the quality ratings domain. Similar to intrusive models, a third order monotonic polynomial mapping of the results is allowed for increased consistency [53]. The fundamental differences among the proposed model architectures are, thus, determined by the type of:

- frame-specific (local) features;
- features supplied to the mapping stage: local or global;
- regression model used in the mapping.

To predict quality in various distortion conditions, a large pool of features is required [43, 26, 27]. Furthermore, as specific types of features are mostly sensitive to specific types of distortions, it is common to use multiple-predictors, the outputs of which are then weighted and combined [43, 27, 54, 55]. Models targeting particular distortion conditions, such as reverberation and dereverberation artifacts, have also been proposed [56].

Vocal tract descriptors [57, 43] and disturbance characteristics such as type and intensity are often used as frame-specific features. Similar to intrusive quality modeling, auditory-domain features are well-represented due to the a-priori compaction of the relevant information that these achieve. Among others, temporal-envelope-based features are used in [58, 27, 56]. Perceptual linear prediction (PLP) coefficients are used in [55]. Features extracted from a time-bark frequency scale spectrogram are used in [59]. High predictive performance can also be achieved without the use of features with auditory motivation [26].

Feature selection is important for reducing over-fitting to the training data and maximizing performance in arbitrary test scenarios. Due to the higher dependence on the training data, this problem is more severe for non-intrusive models. Expert knowledge is used in most cases for a priori selection of a subset of features for quality prediction [43, 27]. Further sparsification can be achieved with signal processing techniques such as $L_1$ norm constraints [60, 61], and penalties on the model complexity such as, e.g., the minimum description length (MDL) [60]. If the model is defined in a Bayesian set-up, sparsity-inducing prior probabilities can be used instead [46].

Regarding the type of features supplied to the mapping stage, the most prominent approach is to first derive global features representing the entire utterance and then map these to the quality domain [43, 26, 62]. Averaging and extreme value analysis are used in [43] to obtain the global features. High-order statistical analysis is used in [26]. Likelihood evaluation on clean speech models of the frame-based features provides the global descriptors in
A prominent example of the approach where frame-based features are first mapped to instantaneous quality ratings and then combined (through time-averaging) is [27].

The sophistication of the regression models used for mapping features to quality ratings varies significantly from model to model. There is, however, a noticeable trend in the direction of using advanced machine learning methods. A hierarchical linear regression is used in [43]. Marginalizing the feature dimensions from a joint feature-plus-rating Gaussian mixture model (GMM) gives the predictive quality score in [26]. Multivariate adaptive regression splines (MARS) are used in [55], support vector regression (SVR) and neural network (NN) perform the mapping in [63] and [27] respectively. A hierarchical Bayesian model captures the heterogeneity of the training data in [54].

It is interesting that despite the discrete nature of the raw quality ratings and the broad use of discrete choice models (DCM) in other disciplines [64], these are not used widely [65, 12] in quality assessment applications. A plausible explanation is the high complexity these incur. To achieve maximum benefit from the information contained in the training data, however, it is necessary to use models that take into consideration the statistical properties of this data. It is, therefore, likely that DCMs will see a wider application to QA in the future.

**1.2.3 Parametric QA**

Parametric QA models predict quality from a small set of parameters that describe the operation state of the communication equipment. These include, among others, packet loss, delay, jitter, codec type and bit-error rate. Notably, most of the afore-mentioned parameters can be obtained from the header of a packet in a packet-based network, and all of them can be computed without the need for decoding the payload. Parametric QA, thus, provides a powerful tool for monitoring quality without significantly affecting the throughput of the communication equipment.

The E-model [44] predicts quality degradation due to codec rate constraints, network delay and distortion co-occurring with the information-bearing signal. It is recommended primarily as a planning tool for use with arbitrary networks. When the type of network is known a priori, it is feasible to limit the number of parameters and learn a mapping to quality ratings using data from subjective listening tests [66].

**1.3 Models for Predicting Intelligibility**

Efforts to develop objective models of speech intelligibility precede work on quality assessment [1]. An intuitive explanation is that intelligibility as a measure is closely-related to the transfer of information. In that sense,
quality is a secondary attribute. Parallels between the development of information theory and intelligibility measures are highlighted, e.g., in [67].

Similar to QA, both intrusive and non-intrusive intelligibility models (IMs) are proposed in the literature. Established QA models [40] have also lately been considered for measurement of intelligibility [68]. Joint modeling of quality and intelligibility for specific distortion conditions has been considered in [69, 56]. Most of the work on IMs is focused on intrusive methods, i.e., methods where both the clean signal and its noisy version or a statistic of the disturbance are available. A review of selected methods, including a classification according to operational principles is available in [68].

Intrusive intelligibility modeling is of particular interest in this thesis as it identifies measures that can be optimized for the parameters of selected speech modifications to enhance intelligibility in adverse conditions. Various criteria can be employed to classify intrusive IMs. Based on the domain where intelligibility is evaluated, there are i) low-level [70, 71, 15], i.e., operating at the signal or signal spectrum level and ii) high level [17], i.e., operating at a level where phonetic information is visible. A different type of classification recognizes models that either i) require the clean speech signal and the disturbance (or a statistic thereof) [15] or ii) the clean speech signal and the noisy signal [72, 73, 74]. In the following, some of the main contributions to the field are reviewed.

1.3.1 Low-level methods

The articulation index (AI) [1, 70] has served an important role in the development of intelligibility measures as it has demonstrated the relative importance of spectral bands in the speech spectrum. It evaluates the proportion of speech audible to the ear from the weighted contributions of individual bands. Due to its limitations, among others its inability to predict reverberation-induced intelligibility degradation, AI has been superseded by newer methods.

The speech transmission index (STI) [75, 71] links intelligibility degradation to reduction in signal modulation. Deviation of the modulation depth, i.e., the ratio of the modulation and the carrier amplitude, from the clean signal reference indicates an increase in the level of the disturbance in the considered frequency range. Variations of STI use either speech-based or noise-based test signals. Different methods for computing STI exist and are reviewed in [72]. Apart from additive noise, STI accounts for non-linear distortions, reverberation and echos.

An evaluation of the performance of AI and STI for conversational [38] and clear speech [76, 77] in noise and reverberation [69] reveals an interesting result. While both approaches predict, in part, intelligibility loss due to signal degradation, neither explains the intelligibility variation due to
speaking style differences. The implication is that low-level methods are insensitive to the intelligibility enhancing effect of modifications that fall outside of the parametric domain where these operate.

The speech intelligibility index (SII) [15] derives from STI and, similarly, accommodates several methods of calculation depending on the number of bands and their distribution in frequency. It uses higher frequency resolution compared to STI and achieves good correlation with subjective results for reverberation, noise and distortion. Late reflections and echos, compression, limiting and non-linear phase modifications are not handled properly by the method. Notably, dynamic range compression has been shown to produce significant intelligibility gain for speech in noise [78, 79].

An extension of SII, the coherence SII (CSII) [73] further accounts for clipping distortions common in hearing aids. The key modification is replacing SNR by signal-to-distortion ratio (SDR), which is computed in terms of a coherence function [80] between the input and the processed signals.

An intelligibility predictor designed on the basis of a sophisticated auditory model [81, 82] is proposed in [83]. The model accounts for masking effects in the auditory periphery and is shown to perform well for additive noise and additive noise in combination with ideal time-frequency segregation (ITFS) [84]. Under ITFS, time-frequency cells where the level of the signal satisfies a certain relation with respect to the noise level are preserved and eliminated otherwise. The relation can be adapted to manipulate the number of preserved TF cells. ITFS, thus, models the behavior of noise-suppression algorithms. As the level of detail in a model increases, so does the computational effort. The benefits of increased accuracy and extended operating range in terms of distortion conditions, thus, come at a price, which may be unattractive for intelligibility enhancing speech modification optimization.

Short-time objective intelligibility (STOI) [74] is another model targeting intelligibility prediction for noisy speech processed by noise-suppression algorithms. STOI performs correlation analysis between the clean and the noisy signals. The high computational efficiency gives this method a practical advantage. A model predicting intelligibility by evaluating the mutual information between long-term clean and noisy signal temporal envelopes is proposed in [85]. On average, its performance is similar to that of STOI.

1.3.2 High-level methods

The glimpsing model [17] is a representative from a small group of models that quantify intelligibility at a level where phonetic cues can be taken into account. The front-end of the model, commonly referred to as glimpse proportion and widely applied for intelligibility enhancing speech modification [35, 86], generates a time-frequency map of speech “visibility” in the presence of the distortion. Missing-data speech recognition is then employed
to measure intelligibility. High complexity and challenges related to using general speech material limit the practical application of the method.

A high-level measure of speech intelligibility, based on the probability of correct recognition, is proposed in [18] (paper C). The measure is tailored for intelligibility enhancing speech modification and validated in an additive noise scenario. The method avoids the need for performing computationally-costly automatic speech recognition (ASR). Good performance is conditional on accurate sound segmentation and the use of a speech model that is representative of the test material.

1.4 Intelligibility Enhancing Speech Modification

The quality and the intelligibility of speech decrease in the presence of additive noise and distortions such as reverberation and speech processing artifacts. Section 1.2 summarized some of the main contributions regarding the development of models that quantify quality degradation, and Section 1.3 focused on the corresponding work from the perspective of intelligibility. Monitoring quality and intelligibility is important for diagnostic purposes. A related problem with strong practical implications is the enhancement of these attributes with the objective of improving QoS.

Speech enhancement [87, 88, 89] focuses on the suppression of noise in a mixture of speech and noise. Such methods achieve effective quality improvement, but do not necessarily benefit intelligibility [90]. When the clean speech signal is known, it is feasible to modify it before it is rendered in a noisy environment, enhancing its intelligibility [91, 92]. Typical application scenarios include public announcement systems, speaker phones and dialog systems. Ongoing effort in the field has produced algorithms that boast significant intelligibility gain without increasing the power of the signal [35, 93].

Speech modification methods are classified according to several criteria. Methods for which the degree and type of the modification adjusts to the changes in the environment are adaptive. Adaptation is typically achieved by optimizing [94] a measure of intelligibility [95, 18, 37, 86]. As such measures are functions of the properties of the speech and the distortion, a change in any of these two factors affects the optimal modification. While non-adaptive methods [78, 79] can achieve very high performance at a low computational cost [34], from a theoretical perspective, the degree of modification at any given time is suboptimal. Furthermore, when the distortion level is low, deviation from the natural signal may degrade the perceived voice quality. Psychoacoustics [96, 97], provide numerous other examples in favor of adaptation, e.g., the dependence of energetic masking on the power spectral density of the noise.

Another classification criterion is related to the message representation level where the modification is applied. Lexical modifications such as, e.g.,
rephrasing [98] are high level modifications. Prosodic modification is observed at a level below lexical modification as the transcription of the message is fixed. Spectral modifications such as band-gain adjustment and spectral envelope manipulation [99, 37, 100, 86], and temporal modification such as dynamic range compression [79] and time scaling [101] are low-level modifications.

A third criterion allows classification with respect to the origin of a speech modification. There are i) human-inspired, \textit{i.e.}, modifications performed consciously or subconsciously by humans producing speech in adverse conditions, ii) rational strategies based, \textit{e.g.}, on expert insights of auditory modeling and cognition [78] and iii) modifications that are the result of functional optimization [20, 102]. A selection of speech modification strategies for enhancing intelligibility are reviewed in Section 1.4.1.

1.4.1 Speech Modification Strategies

Deliverable D1.1 of the Listening Talker (LISTA) project [7] presents a list of forty-four candidate speech modifications. While this list can be extended further, it includes the modification strategies used in the majority of the proposed intelligibility enhancement systems. The effect of some of the listed modifications on perception in noisy environments is reviewed in [34, 93]. Some human strategies such as pitch modification, vowel space expansion and speaking rate reduction do not improve intelligibility consistently when applied to natural speech. This outcome suggests that such modifications may have an auxiliary role or be the result of physical limitations in the speech production mechanism. In the following we review some of the prominent methods in the field.

Lexical speech modification is a high-level approach and consist of: i) repetition to provide additional cues and ii) rephrasing to increase correct recognition probability as a result of better noise robustness and higher predictability. While repetition is inefficient and does not facilitate intelligibility optimization, rephrasing provides an intuitive and attractive platform for modification. High-level modification requires high-level measures and entails higher complexity and the dependence on additional information. A practical rephrasing approach focused on the acoustic characteristics and based on optimizing a measure derived from the probability of correct recognition [98] reports consistent intelligibility gain over a range of test conditions. Further sophistication, such as modeling the predictability of different formulations is expected to improve performance.

Proceeding down the message representation hierarchy, the effect of prosody is visible, \textit{e.g.}, in the intelligibility gain that clear speech has over natural speech [76, 92]. One likely influence is improved word segmentation. Optimizing a measure based on the probability for correct recognition, for the gain factors of the individual phonemes in a word, under a power
preservation constraint, is considered in [18] (paper C). Other candidate modifications include changing the relative duration of phonetic units and shortening units that are more sensitive to energetic masking in favor of more robust units.

Knowledge of the phonetic representation of a message is not required in low-level modifications. These can be subdivided into spectral, temporal and joint spectro-temporal modifications. Spectro-temporal energy redistribution is considered in [99] where the glimpse proportion is optimized. Use of a genetic algorithm to perform the optimization makes this method interesting primarily from a theoretical perspective. A low-complexity approach with high intelligibility gain that performs spectro-temporal energy redistribution by optimizing a perceptual distortion measure is presented in [103, 100]. This method is used as a reference in papers C and D.

Spectral shaping is employed in [104, 37] where the optimal speech filter is obtained by optimizing an SII-based measure. Low complexity, and high intelligibility gain, e.g., [34] make these approaches particularly suitable for application in mobile telephony. [37] is used as the reference method in paper E. Spectral shaping is also employed in [79], where the degree of modification is not adaptive. A related approach using the combination of fixed spectral shaping with auditory motivation and dynamic range compression (DRC) established the ceiling of the performance gain in a large-scale subjective evaluation of speech modification systems [34]. Adaptive post-filtering emulating spectral effects in Lombard speech is proposed in [105]. Also inspired by Lombard-speech, [106] focuses on spectral shaping and vowel space expansion.

Temporal modifications, such as time-scaling [107], pause manipulation and DRC are used in the context of intelligibility enhancing modification [108, 79, 109] on their own or in combination with other modifications. DRC, which emphasizes low-power at the expense of high-power (or additional power) acoustic events is a particularly successful strategy. One of the early works revealing its effectiveness for speech intelligibility enhancement is [78].

Methods that perform functional rather than function parameter optimization, are a recent addition to the range of proposed speech modification strategies. In the context of multi-zone speech enhancement [110, 102], a modification framework is proposed to minimize cross-talk among a number of target rendering zones. The particular modification depends on the choice of distortion measure and is derived under the assumption of an affine signal model. A modification strategy to improve intelligibility, in additive noise, by spectral dynamics recovery is proposed in [20] (paper E). Calculus of variations is used to optimize a distortion measure that quantifies the deviation of the spectral dynamics of noisy speech from the dynamics of clean speech. Noise and speech adaptive optimal compressors can be obtained as particular solutions to the optimization problem.
2 Thesis Directions

Two research directions are explored in this thesis. The strong dependence of the performance of non-intrusive QA models on the amount and spread of the training data motivate focusing on the efficient use of this data. Modeling uncertainty due to individual variations, feature selection for reduced over-fitting and enhancing predictive performance by conditioning a quality prediction on dependent auxiliary random variables are considered. The second direction is speech modification for enhanced intelligibility in noisy environments. The primary application scenario is that of additive noise.

An introduction to the work on QA is provided in Section 2.1. Intelligibility-enhancing speech modification is discussed in Section 2.2. The contributions of the thesis are summarized in Section 2.3, followed by conclusions and an outline of future work in Section 2.4.

2.1 Quality Assessment for Limited Data

Limited training data is a common problem in the development of non-intrusive QA models as their performance exhibits a strong dependence on the amount and spread of this data. It is possible to obtain more data either through additional subjective tests or by pooling together existing databases. Subjective tests, however, are costly and time-consuming, and data pooling opens up the issue of revealing proprietary information related to the evaluation of novel speech processing algorithms and communication technology. There is, thus, a strong motivation to use efficiently the available data.

Efficient use of the data is largely associated with employing techniques that reduce over-fitting. This makes feature-selection a critical stage in the design of a quality assessment model. Modeling additional dimensions, such as uncertainty due to individual variations among test subjects, can also be seen as an enhancement in the predictive performance as it augments the relevant information. This particular dimension is of interest as its effect on the reliability of the mean prediction is typically ignored in the design of QA models. Probabilistic regression models accommodate the variance of a prediction in a natural way [47, 111]. Considering that the variance of bounded support random variables is a function of their mean [112, 113, 114], it is important to use statistical models that capture this dependence. Related contributions are introduced in Section 2.1.1.

Exploiting the correlation between subjective scores from data sets that cannot be pooled together, presents another possibility for enhancing the performance of a quality predictor when the amount of training data is limited. This idea parallels the notion of using multiple experts in machine learning [115]. Augmenting a wide-band (WB) quality prediction with its narrow-band (NB) counterpart provides the target application in paper B.
This is a practically-relevant scenario as a result of the gradual increase of sampling frequencies in speech and audio processing algorithms, following the demand for high-fidelity sound and the availability of larger communication bandwidths. Further details are provided in Section 2.1.2.

2.1.1 Statistical Models for a bounded-support random variable

Quality ratings are obtained using bounded support scales, affecting the properties of the random variable. Averaging multiple ACR quality ratings over a number of test participants, produces MOS, which is a pseudo-continuous random variable constrained to the support range of ACR. Two continuous distribution models are considered in this section as a platform for modeling the variation of both the mean and the variance of quality predictions in a given feature space.

Doubly-bounded continuous distributions are not abundant in the statistics literature. The two most prominent candidates are the Beta distribution [112], a member of the Exponential family, and the Logitnormal distribution [116, 114]. Adopting the Beta distribution allows for processing the data in the original domain but opens convexity and convergence-related issues when fitting the model. The Logitnormal distribution, on the other hand, offers the flexibility of working with Normally-distributed variables in a transform domain. Computing the moments of the distribution in the original domain requires numerical integration, which is computationally-demanding.

The PDF of the Beta distribution is:

$$p_Y(y|\mu, \phi) = \frac{1}{B(\mu \phi, \phi - \mu \phi)} y^{\mu \phi - 1} (1 - y)^{(1-\mu)\phi - 1}, \ y \in [0, 1]$$  \hspace{1cm} (1)

where $\mu$ is the mean, $\phi$ is the precision and $B(\cdot, \cdot)$ is the Beta function. Parameterizing $\mu$ and $\phi$ in terms of features extracted from the signal and maximum-likelihood (ML) parameter estimation is considered in [13] (paper A). Fitting the model to the MOS data gives a simple and effective predictor that compares favorably to the state-of-art in the field in terms of conventional figures-of-merit such as Pearson’s correlation coefficient and root-mean squared (RMS) error. Fitting the model to the ACR ratings using a rich parametrization for $\phi$ indicates the existence of variance patterns in the feature space of the training data and opens the question of whether these can be predicted.

Good predictive performance is closely-related to avoiding over-fitting to the training data. Methods proposed to attain this goal optimize jointly the data fit and an additional regularization term reflecting the model complexity [60, 117] or, in the context of Bayesian inference [46], employ sparsity-inducing prior probabilities. While such methodologies are
generally-applicable, the Beta distribution does not facilitate their implementation. Instead, feature selection in the parametrization of the mean is determined by side experiments, leading to a sub-optimal solution.

Using the Logitnormal in place of the Beta distribution offers an advantage with respect to model reduction. Applying the logit transform:

\[ \tilde{y} = \log \frac{y}{1 - y} \]  

(2)

to a random variable that follows a Logitnormal distribution produces a Normally-distributed random variable. The PDF of the Logitnormal distribution is:

\[ p_Y(y|\tilde{\mu}, \tilde{\sigma}) = \frac{1}{\tilde{\sigma}\sqrt{2\pi}} e^{-\frac{(\log\frac{y}{1-y} - \tilde{\mu})^2}{2\tilde{\sigma}^2}} \frac{1}{y(1-y)}, \quad y \in (0, 1) \]  

(3)

where \( \tilde{\mu} \) and \( \tilde{\sigma} \) are the mean and the standard deviation of the Normal variable \( \tilde{Y} \).

To introduce dependence on the signal, \( \tilde{\mu} \) and \( \tilde{\sigma} \) are parameterized as:

\[ \tilde{\mu} = c^T x \]  

(4)

\[ \tilde{\sigma} = e^{d^T x} \]  

(5)

where \( c \) and \( d \) are model coefficients and \( x \) are global signal features. Inference for the model coefficients is readily performed through ML estimation. Imposing \( L_1 \) norm constraints of the form:

\[ \sum_k |c_k| \leq C \]  

(6)

\[ \sum_l |d_l| \leq D, \]  

(7)

where the magnitude of \( C \) and \( D \) influence model sparsity, provides a convex framework for automated feature selection. Cross-validation for maximum predictive performance is used to determine the optimal constraints.

Fitting of the Logitnormal model to the ACR data from [28] is performed using a recursive procedure where the estimates for \( c \) and \( d \) are conditioned on each other. Convergence is guaranteed by the convexity of the constrained objective. Automated feature selection trims all but the constant-feature term in \( d \) leading to constant variance in the Normal domain. This result suggests that the effect of factors such as trend significance, model simplicity and quantization noise introduced by the categorical nature of the subjective rating scale needs to be taken carefully into consideration.

Fitting the model to the corresponding MOS data produced a different outcome. In this case, data-dependent variance patterns were successfully identified. Conditioning the estimate of \( c \) on the estimate of \( d \) resulted in consistent but limited improvement in predictive performance as measured by Pearson’s correlation coefficient and the RMS error.
2.1.2 Prediction Augmentation with Correlated Data

Prior information about a random variable and the existence of dependence among random variables, provide the possibility for reducing uncertainty and improving predictive performance. Given that non-intrusive QA is approached as a statistical modeling problem, exploiting fully such relations is advantageous when the training data is scarce.

Bayes’ theorem [118] provides the means to include prior knowledge when performing inference for a random variable:

$$ p(y|x) = \frac{p(x|y)p(y)}{\int_y p(x|y)p(y) \, dy}, \quad (8) $$

where $p(y)$ is the prior distribution. If the random variable $Y$ is jointly distributed with some other random variable $Z$, the uncertainty about $Y$ can be reduced by conditioning on $Z$. The conditional distribution can be obtained from the joint distribution by applying the theorem on compound probabilities (chain rule) [119] and identifying the desired expression:

$$ p(y, z|x) = p(y|z, x)p(z|x). \quad (9) $$

Alternatively, applying Bayes rule directly to the conditional distribution gives:

$$ p(y|z, x) = \frac{p(x|z, y)p(z, y)}{p(z|x)p(x)}. \quad (10) $$

A method exploiting the dependence between the target quality variable and another random variable that characterizes quality in related conditions is proposed in [14] (paper B). In favor of tractable derivations, the quality ratings are modeled with Gaussian processes (GP) [120, 47]. Regression models based on GP are referred to as non-parametric because there is no explicit specification of the relation between the dependent and the independent variables. The choice of a function describing the data covariance represents the model designer’s understanding of the data. Application of GP to regression problems provides the means for uncovering complex data relationships with few modeling assumptions.

An application scenario used to validate the proposed approach is conditioning a WB quality prediction on its NB counterpart. A probabilistic NB quality prediction is obtained from a model trained on data that cannot be pooled together with the data used for training the main predictor due to difference in the sampling frequencies. This approach allows for reusing legacy QA data to improve the performance of new measures, in particular, when training data is scarce. Propagating the uncertainty of the NB prediction regulates its influence on the WB estimate and prevents performance degradation. Simulation results indicate improvement in predictive performance emphasizing the potential for real-world application. The adopted
methodology is easily generalized to accommodate the scenario of combining predictions from multiple pre-trained models.

2.2 Intelligibility Enhancing Speech Modification

The mutual information between a transmitted and the corresponding received messages decreases with an increase in the power of the environment noise. With respect to baseband speech communication this result means reduction in intelligibility and quality. Increasing the speech power offers a trivial solution by improving SNR. When the noise power is already high, speech power increase may lead to distortions in the communication equipment, listener discomfort, and even hearing impairment. Speech modification before presentation in a noisy environment offers an alternative solution and has been shown to produce significant intelligibility gain. This section introduces the thesis research on the topic of intelligibility enhancing speech modification.

The communication process has an hierarchical structure [121, 8] and can be described mathematically as a first-order Markov process [18, 122]. A representative illustration is given in Figure 1 where $M$, $U$ and $A$ denote message, utterance and acoustic features respectively. The indexes $S$ and $L$ identify the speaker and the listener. A bold arrow identifies the link where the communication channel affects intelligibility. Effectively, communication begins with the formulation of message $M_S$. At a lower level, this message is mapped into a sequence of phonemes represented by $U_S$. The phoneme sequence is then projected onto the physical layer where it is described by the acoustic features $A_S$. The communication channel may introduce distortion leading to a different acoustic representation $A_L$ on the listener side. The acoustic features are interpreted as a sequence of phonemes $U_L$ that provide the basis for the decoded message $M_L$.

A measure of intelligibility can be formulated at any of the hierarchical levels of the communication process by evaluating the similarity between the speaker and the listener side representations. Considering that the objective of communication is the exchange of information, a suitable measure of

\[ \text{Intelligibility} = \text{Similarity}(M_S, M_L) \]

Figure 1: A Markov model representation of the speech communication process.
similarity is the mutual information \([123]\). Ideally, the similarity measure is evaluated at the message level \([122]\):

\[
I(M_L; M_S) = \sum_{M_L, M_S} p_{LS}(M_L, M_S) \log \frac{p_{LS}(M_L|M_S)}{p_L(M_L)},
\]

(11)

where \(p_{LS}\) and \(p_{L|S}\) are the joint and the conditional probabilities of the intended and the perceived messages respectively. Separating the contributions of the correctly decoded message from all incorrect interpretations gives:

\[
I(M_L; M_S) = \sum_{M_S} p_{LS}(M_S, M_S) \log \frac{p_{LS}(M_S|M_S)}{p_L(M_S)} + \sum_{M_L \neq M_S, M_S} p_{LS}(M_L, M_S) \log \frac{p_{LS}(M_L|M_S)}{p_L(M_L)}.
\]

(12)

As the probability of correct recognition \(p_{L|S}(M_S|M_S)\) approaches unity, the first term approximates the entropy of the intended message, and the second term approaches zero. This relation motivates a simpler approach to intelligibility enhancement based on the maximization of \(E_S\{\log (p_{L|S}(M_S|M_S))\}\), instead of the less manageable \(I(M_L; M_S)\).

At the current state of technology, it is not possible to optimize a message-level measure without involving the subject. As this is inefficient for practical applications, it is necessary to consider optimization at the lower levels of the communication hierarchy. Most optimization-based methods adopt measures defined at the level of acoustic features. The relation between mutual information and SII \([15]\), one of the most prominent intelligibility measures, is identified in \([124]\). A measure defined at the level of the phonetic contents and used for intelligibility enhancement is proposed in \([18]\) (paper C).

It is important to note that a measure of intelligibility defined at a certain level is responsive to modifications that take place at the same or any of the lower levels. Thus, optimizing SII for prosodic or rephrasing-based modifications is not expected to provide relevant feedback. A consequence is that working at a higher abstraction level allows for a wider range of visible modification strategies. The computational complexity and the amount of required information for applying the measure increase as well.

The phoneme-level measure of \([18]\) (paper C) is introduced in 2.2.1. Addressing complexity and robustness issues, associated with high-level measures, a projection onto the lower (acoustic) level is discussed in 2.2.2.

### 2.2.1 Probability of Correct Phoneme Recognition

The first level down the communication process hierarchy, where an objective measure can be defined and optimized in an autonomous way, is
that of the utterance, i.e., the sequence of phonemes that encode the message. In terms of the probability of correct recognition, this measure is $E_S\{\log (p_{L|S}(U_S|U_S))\}$. Using the ergodicity [125] of the communication process, the expectation over the intended message space can be replaced by time-averaging.

A different starting point is used in [18] (paper C) to derive a practical high-level objective measure for the phonetic sequence. Instead of the mutual information, the probability of correct recognition comes out as a result of applying a hit-or-miss distortion criterion to the decoded phonetic transcription. The difference between the two measures sits in the absence of the $\log(.)$ function under the expectation operator in [18]. Asymptotically, i.e., with an increase in the time duration of the acoustic message representation, the two measures are equivalent. It is important to note, however, that time constraints apply when modifying speech for enhanced intelligibility. Factors such as, e.g., round-trip delay and relevance of the noise statistics determine the length of the modification window.

Physically meaningful approximations motivated with the low uncertainty in the clean speech features and the stationarity of the noise simplify the expression for the selected distortion criterion. Following an optimality preserving transformation and equivalent transformations based on the application of Bayes rule, the following practical measure is derived in paper C:

$$\mathcal{O} = \log (p(\hat{a}_L(C^*)|u_S,V)) + \log (p(u_S|V))$$

$$- \log \left(\sum_{u_L \neq u_S} p(\hat{a}_L(C^*)|u_L,V)p(u_L|V)\right),$$

where $\hat{a}_L$ is an estimate of the listener-side acoustic features, $V$ is a mathematical model of speech and $C^*$ are generic modification parameters. The first term is the likelihood of the distorted features given the correct acoustic models and the third term discriminates against all alternative phonetic sequences. The second term reflects the prior probability of the formulation used to encode the message and is independent of $C^*$ given that formulation modification is not considered.

Only the first term is optimized in [18] due to the complexity incurred by the computation of the third term. Two cascaded modifications are considered. One of these affects the prosody by distributing word energy among the word phonemes while the other distributes spectral energy in an auditory filter-bank domain.

Results from evaluating the effect of selective discrimination, i.e., approximating the third term with a partial sum [34], indicate that the theoretically-optimal measure is also better in practice. The study reiterates the importance of accurate sound segmentation information and a representative speech model to the performance of the method.


2.2.2 Spectral Dynamics and Speech Intelligibility

Use of high-level intelligibility measures extends the range of possible signal modifications and allows for optimization at a level closer to that of the message. As noted in the beginning of Section 2.2, enhancing the similarity between the intended and the received message is the ultimate goal of intelligibility enhancing speech modification.

Along with its advantages, however, use of high-level measures implies higher complexity and requires a significant amount of additional information such as transcription, sound segmentation and an acoustic model of speech. It is attractive, therefore, to consider a solution that can, in part, preserve the advantages of high-level measures while increasing robustness and reducing the amount of required information. One solution is to map a high-level measure into the feature space. Increasing the similarity of the modified noisy and the natural speech features, through signal modification, is expected to also increase the probability of correct recognition.

The neuroscience literature emphasizes the importance of the temporal evolution of the speech signal to intelligibility [11, 126, 127]. In accord with these findings, the focus here is on increasing the similarity between the noisy modified and the natural speech dynamic features as defined in the context of ASR [128, 129]. The problem can be approached from the perspective of dynamics preservation, i.e., restoration subject to a lower threshold on the power level, or dynamics recovery where a distortion measure is optimized under a power constraint.

Dynamics preservation Initially, efforts were concentrated on preserving the spectral dynamics [19] (paper E). Using the result that frame-wise differences of MFCCs [130] are preserved when the differences of the log band-powers of the corresponding frames are reinstated, the problem reduces to enforcing the equality:

\[
\frac{b_{i,j+l}m_i^T\Phi_{x_{j+l}} + m_i^T\Phi_{n_{j+l}}}{b_{i,j}m_i^T\Phi_{x_j} + m_i^T\Phi_{n_j}} = \frac{c_i m_i^T\Phi_{x_{j+l}}}{c_i m_i^T\Phi_{x_j}}, \quad i = 1, \ldots, I, \tag{14}
\]

where \(i\) is a spectral band index, \(j\) is a frame index, \(\Phi_{n_j}\) and \(\Phi_{x_j}\) are the periodograms of the noise and the speech in frame \(j\) respectively, \(m_i\) is the Mel filter response in band \(i\), \(b_{i,j}\) is a gain modification parameter and \(c_i\) is a free parameter. As there are two unknowns \((b_{i,j+l} \geq 0 \text{ and } b_{i,j} \geq 0)\) and a single equation, there is an infinite number of solutions for a fixed value of \(c_i\). A particular solution is obtained by enforcing independently the equalities of the numerators and the denominators from (7):

\[
b_{i,j+l} = c_i - \frac{m_i^T\Phi_{n_{j+l}}}{m_i^T\Phi_{x_{j+l}}}, \quad \forall l, \tag{15}
\]
where the value of $c_i$ is related to the power threshold above which preservation can be achieved. Importantly, untying the modification gains reduces system delay and allows efficient computation. The optimal parameters $c_i$, $i = 1, \cdots, I$ are derived by minimizing the average output power under a constraint on the average dynamics preservation level across all bands. It is also possible to reverse the places of the objective and the constraint and focus on maximizing the dynamics preservation level under a constraint on the output power.

Subjective evaluation of the method validated its capacity for increasing intelligibility in noisy environments. The power efficiency of the method, however, is low. In particular, since dynamics preservation is progressively performed in a direction from high to low band-powers a lot of energy is allocated to signal time-frequency cells that, likely, already are intelligible. Allowing for distortion in the reconstructed spectral dynamics opens the possibility to address this issue.

**Dynamics recovery** An approach to optimal dynamics recovery is presented in detail in [20] (paper D). To facilitate the derivation, the distortion measure is defined in continuous time domain. Denoting the modified speech signal by $y$ and considering an additive noise model and a single-band scenario, the dynamics are recovered perfectly when:

$$\frac{d \log (y(x) + n)}{dt} \frac{dt}{d \log (x)} = 1,$$

where the dependence of $x$ and $y$ on the time $t$ is not indicated explicitly. Defining a weighted squared-error distortion measure and adding a power-related constraint produces a distortion criterion of the form:

$$\eta_1 = \int_0^\beta \left( \frac{1}{x} (y(x) + n - x y'(x))^2 + \lambda g(y, x, n) \right) f_X(x) \, dx,$$

where $f_X(x)$ is a statistical model of the band-powers, $\lambda$ is a Lagrange multiplier, $g(y, x, n)$ is a function of the output power, and $\alpha$ and $\beta$ determine the optimality range.

The solution to (17) for a gain (linear) constraint and Pareto distributed $x$ [131, 132] is derived in [20] (paper E) using calculus of variations [133]. The single-band solution is then used as the basis for a multi-band system, where power is optimally distributed among frequency bands using a second-order distortion criterion that maximizes the mean recovery level.

In the following, the result from the optimization of criterion (17) for a second-order penalty on the output power and an adjusted weighting of the distortion function is summarized. This case is interesting as it provides a solution in terms of exponential functions, unlike the case with a
linear penalty where the solution is polynomial. Closed-form solution, however, could only be derived for two particular values \{\frac{1}{2}, \frac{3}{2}\} of the shape parameter of the Pareto distribution.

The integrand of the new criterion for a squared penalty on the total power reaching the ear is:

\[
L(x, y, y') = \left( \frac{n}{x} (y + n - xy')^2 + \lambda (y + n)^2 \right) A x^{-(1+b)},
\]

(18)

where \(y' = \frac{dy(x)}{dx}\), \(A\) is a constant and \(b\) is the shape parameter of the Pareto distribution. The dependence of \(y\) on \(x\) is omitted for brevity of notation. The adjustment of the distortion function weight shows in the additional factor \(n\) which counteracts the increased influence of the noise power in the penalty term. A closed-form solution can be derived without this adjustment as well, however, it is not of practical significance.

The fundamental lemma in calculus of variations [133] identifies a necessary condition for an optimizer in the form of the Euler-Lagrange equation. Applying the lemma to criterion (18) for \(b = \frac{1}{2}\) and simplifying gives the following linear non-homogeneous ODE [134]:

\[
y'' - \frac{1}{2x} y' + \left( \frac{1}{2x^2} - \frac{\lambda}{nx} \right) y + \frac{n}{2x^2} - \lambda \frac{1}{x} = 0.
\]

(19)

Using standard techniques [135, 136], the solution to (19) is obtained of the form:

\[
y = c_1 x^{\frac{1}{2} e^{\sqrt{\frac{\lambda}{n} x}}} - c_2 \frac{1}{\sqrt[4]{\lambda}} x^{\frac{1}{4} e^{-2\sqrt{\frac{\lambda}{n} x}}} - n,
\]

(20)

where \(c_1\) and \(c_2\) are constants. To establish a particular solution these constants are determined by formulating physically meaningful initial and final conditions. One possible set of conditions would ensure that the function \(y(x)\) achieves two particular points in the range of interest. Another possibility is to achieve one particular point and ensure specific behavior at another point by fixing the value of the derivative.

Relation (20) between the original and the modified (for improved intelligibility in noise) output signal powers has potential for application in practice. Suitable initial and final conditions ensure two desirable properties: i) output power increases with an increase in the noise power and ii) an increase in the value of the Lagrange multiplier \(\lambda\) decreases the output power. Since a closed-form solution only exists for particular band-power statistics, it would be of interest to explore different approximations that would facilitate general application of the considered criterion.

### 2.3 Specific Contributions

Five papers are included in the second part of this thesis. Papers A and B address problems in non-intrusive QA. Papers C, D and E focus on in-
telligibility enhancing speech modification. The individual contributions of each of the papers and the candidate are summarized in the following.

Paper A employs a statistical framework suited for modeling bounded-support variables and applies it to non-intrusive QA. The predictive performance of the model, for the mean of the quality prediction, compares favorably to the state-of-art in the field while using a limited training database. A signal-feature parametrization of the variance of the random variable provides a platform for modeling patterns due to individual variations in the feature space of the training data. The contributions of the candidate include problem formulation in discussion with supervisor, proposing the methodology and deriving the solution.

Paper B addresses the problem of enhancing the performance of a quality predictor when the training data is scarce but a dependent random variable or data that cannot be pooled together with the main training data are also available. The adopted GP framework facilitates tractability in the model derivation. An auxiliary, probabilistic quality prediction is seamlessly included in the feature set of the main predictor. By taking into account the uncertainty of the auxiliary quality prediction the performance of the main predictor is enhanced further by avoiding biasing to unreliable auxiliary predictions. The contributions of the candidate include problem formulation in discussion with supervisor, proposing the methodology and deriving the solution.

Paper C extends an hierarchical perspective of the speech communication process and uses it as a platform for comparison and analysis of intelligibility-enhancing speech modification methods. Further, it introduces a measure for assessing intelligibility degradation due to signal distortion, which is based on the probability of correct recognition of the phonetic sequence encoding a message. The measure is used for optimizing intelligibility enhancing speech modifications for speech in noise. The high abstraction level at which the measure is defined separates modification type from measure specifics and facilitates the joint application of a range of signal modifications. In particular, prosody-affecting phoneme gain adjustment and spectral-energy redistribution are optimized in cascade at a time scale determined by the word duration. Significant intelligibility gain is measured when accurate sound segmentation and a representative speech model are available. The contributions of the candidate include developing the concept, initially identified by the supervisor, for using automatic speech recognition (ASR) techniques in speech intelligibility enhancement, designing and validating a particular method instantiation.

Papers D and E propose and validate two variants of a framework for intelligibility enhancing speech modification, for speech in noise, which effectively maps a high-level measure onto the feature space. Retaining asymptotically the conceptual advantage of high-level measures, the proposed methods waive the requirement for a message transcription and sound seg-
mentation information. Focusing on the dynamic features, optimal relations between input (original) and output (modified) signal power are derived and used in the design of multi-band systems. Paper D approaches the problem from the perspective of preserving the spectral dynamics of the speech band-powers down to a band-specific power threshold. Paper E explicitly allows for distortion in the recovered dynamics at all power levels and improves significantly power efficiency. The single-band solution identified in paper E generalizes the application of DRC in intelligibility enhancement by deriving compressors adaptive to the noise and the speech statistics. High intelligibility gain is achieved at a very low computational complexity. The contributions of the candidate, regarding paper D, include the design of the single-band theoretical framework in cooperation with the supervisor, developing a multi-band extension, designing and validating a method instantiation. With respect to paper E, the contributions of the candidate include developing the methodology outlined at a high level by the supervisor, deriving two particular single-band solutions, developing a multi-band approach, designing and validating a method instantiation.

2.4 Conclusions and Future Work

The methods and the physical instantiations described in the papers enclosed in Part II of this thesis, as well as the supporting publications, all propose functional solutions that can be applied directly in practice. Room for improvement as well as alternative methodologies offering promise of better performance exist as well. Specific ideas regarding the individual publications are listed in this section.

The Beta regression model described in paper A was compared, cf. Section 2.1.1, to an alternative model based on the Logitnormal distribution. Normally distributed random variables in transform domain facilitate automated feature selection and, in turn, reduce the risk of over-fitting the training data. The difficulty in predicting variance patterns due to individual differences among subjects and the limited gain when conditioning the estimate of the parameters of the mean on the estimated parameters of the variance for average ratings may also be indicative of a poor modeling assumption. In particular, these results raise the question of the suitability of a single-level model when multiple distortion types and levels are represented in the training data. As discussed earlier, a number of the established QA models employ layered architectures with multiple predictors the contributions of which are weighted together [43, 27]. A similar sophistication is likely to benefit variance modeling as well.

With regards to paper B, a quality predictor that takes advantage of an auxiliary stochastic prediction is generally useful. A particularly attractive feature of the model is that irrelevant auxiliary predictions have little influence on the target prediction. The complexity of the method, however,
increases with the amount of training data, which is a general problem with GP [47, 65]. Approximations have been proposed based on using: i) a representative subset (global) [137, 138] of training points, ii) a neighborhood subset (local) [139] of training points and iii) combinations thereof [139]. The effect of these approximations on the performance of the augmented quality predictor requires assessment. Going beyond the particular objective of the paper, it is of interest to extend the range of training data protocols that can be combined. In particular, combining ACR, CCR and DCR data to provide predictions on any of these scales is the ultimate goal. In theory, this can be achieved by discrete choice models (DCM) [64].

The separability of the quality perception mechanism into two stages, absolute utility and response formulation holds the key to combining arbitrary training data types. Knowledge of the utility model is refined using any data. The response formulation, e.g., comparison or absolute rating elicitation, in theory, only requires information regarding the target scale. The challenge with using DCMs is the high computational complexity related to the need for performing stochastic integration when computing the predictive distribution [64, 140, 141]. Use of a GP-based approximation reduces, in part, the computational cost. A closed-form approximation to the predictive distribution of the ordered probit [64] is under development and provides indication of significant decrease in complexity with minor performance degradation.

The intelligibility enhancement framework from paper C has both theoretical and practical significance. The requirement for transcription of the message can be waived by the use of a recognizer. Various experiments were conducted in this direction and the main conclusion is that while performance loss occurs, it is a minor issue compared to the increase in complexity. Another aspect in need of attention is the development of modification strategies that take advantage of knowing the particular sounds to which modification is applied. Thus, instead of using a general band-gain adjustment procedure, one can, e.g., perform formant-aware spectral modification for voiced sounds. The expected outcome is reduced distortion, enhanced intelligibility and faster convergence during optimization.

Spectral dynamics recovery, proposed in paper E, provides a computationally efficient method for achieving significant intelligibility gain in various noise environments. Biasing of the optimal solution was performed to reduce processing artifacts that limit the intelligibility gain. Removing the need for biasing offers a conceptual advantage as it cleans up the modeling framework. The procedure is not unmotivated because, effectively, it acts as a rudimentary constraint on signal quality degradation. Two directions can be identified in achieving the desired result with other means. A likely solution is the use of a joint density in the multi-band optimization problem. This would penalize the allocation of energy to spectral bands that lead to sound distortion but also complicates the mathematical analysis. Another
approach is identified by recognizing that the intelligibility reducing artifacts occur primarily in the high frequency range. It is the high-frequency bands that exhibit stronger concentration at low power levels and consequently lead to steeper increase of the function describing the input-output power relation. Reducing the sensitivity of the measure to the shape parameter of the speech band-power statistics is likely to mitigate the problem.

An interesting research problem regarding the use of high-level measures [18, 122] is the joint application of speech modification at all levels of the communication hierarchy summarized in the beginning of Section 2.2. A vision of a future speech modification system includes the possibility for applying signal-level modifications to all possible formulations of a message to determine the modified speech utterance that achieves the highest intelligibility gain.

References


