SPEECH ASSESSMENT AND CHARACTERIZATION FOR LAW ENFORCEMENT APPLICATIONS

DUSHYANT SHARMA

JULY 2012

A Thesis submitted in fulfillment of requirements for the degree of
Doctor of Philosophy of Imperial College London

Communications and Signal Processing Group
Department of Electrical and Electronic Engineering
Imperial College London

Abstract

Speech signals acquired, transmitted or stored in non-ideal conditions are often degraded by one or more effects including, for example, additive noise. These degradations alter the signal properties in a manner that deteriorates the intelligibility or quality of the speech signal. In the law enforcement context such degradations are commonplace due to the limitations in the audio collection methodology, which is often required to be covert. In severe degradation conditions, the acquired signal may become unintelligible, losing its value in an investigation and in less severe conditions, a loss in signal quality may be encountered, which can lead to higher transcription time and cost.

This thesis proposes a non-intrusive speech assessment framework from which algorithms for speech quality and intelligibility assessment are derived, to guide the collection and transcription of law enforcement audio. These methods are trained on a large database labelled using intrusive techniques (whose performance is verified with subjective scores) and shown to perform favorably when compared with existing non-intrusive techniques. Additionally, a non-intrusive CODEC identification and verification algorithm is developed which can identify a CODEC with an accuracy of 96.8 % and detect the presence of a CODEC with an accuracy higher than 97 % in the presence of additive noise.

Finally, the speech description taxonomy framework is developed, with the aim of characterizing various aspects of a degraded speech signal, including the mechanism that results in a signal with particular characteristics, the vocabulary that can be used to describe those degradations and the measurable signal properties that can characterize the degradations. The taxonomy is implemented as a relational database that facilitates the modeling of the relationships between various attributes of a signal and promises to be a useful tool for training and guiding audio analysts.

Declaration of Originality

I declare that this thesis is the outcome of original research conducted by myself under supervision from Patrick Naylor and Mike Brookes. Any work that has been previously published and included in this thesis has been fully acknowledged in accordance with the standard referencing practices of this discipline. Moreover, no part of the research presented in this thesis has been submitted for any degree at any other institution.

Dedicated to H.H. Shri Mataji Nirmala Devi

Acknowledgments

.

I begin by thanking Patrick Naylor and Mike Brookes for supervising this thesis, for sharing their expertise and lending their support in the course of this PhD, without which this would be a lesser thesis. It has indeed been an honor to work with you! Also, the CLEAR team members: Nick, Gaston and Mark have been an inspiration for me and their collaboration has made me a better researcher for which I am very grateful. A note of thanks goes to my CSP colleagues and conference buddies: Sira Gonzalez, Yu Wang, Jason Filos, Daniel Jarret, Mark Thomas and James Pearson. Thanks to Hojjat Akhondi for the brilliant ideas over the many coffee breaks, that could have made us millionaires (but unfortunately didn't)!

I would like to thank my parents, Shruti and Vijay for their unending patience and love and my sister Saumya in whose lovely company I spent the last part of my PhD. It is difficult to imagine myself getting this far without all your love and prayers! Also, I would like to thank my grand parents for all their love and encouragement throughout my life. I thank Peter Yeboah, who is an example for me and the SYUK collective for their divine company and all the support throughout my London life and finally I thank Shri Mataji for the endless blessings that have made me what I am today. May this knowledge that I have been blessed with be used for the betterment of society.

-Dushyant Sharma

Contents

List of Figures

List of Tables

List of Abbreviations

ACR Absolute Category Rating 51

AI Articulation index 103

AIR Acoustic Impulse Response 136

AMR Adaptive Multi Rate. A narrow-band speech CODEC 123

ANSI American National Standard Institute 71

ASR Automatic Speech Recognition 138

BASIE Bayesian Adaptive Speech Intelligibility Estimation 103

CART Classification and Regression Trees 31

CELP Code Excited Linear Prediction 123

CCR Comparison Category Rating 69

CODEC Coder-Decoder 27

DAM Diagnostic Acceptabiliity Measure 139

DCR Degredation Category Rating 69

dBOV dB Overload . Point where clipping occurs 77

dBSPL dB Sound Pressure Level. Signal level in terms of sound pressure 144

EGG Electroglottograph. A device for measuring the vibration of the vocal cords. 47

FFT Fast Fourier Transform 47

GCI Glottal Closure Instants 47

GOI Glottal Opening Instants 47

GMM Gaussian Mixture Model. An approximation to an arbitrary probability density function that consists of a weighted sum of Gaussian distributions 43

GSM-FR GSM Full Rate. A GSM CODEC operating at 13 kbit/s 123

IRS Intermediate Reference System 74

iSNR Importance weighted Signal to Noise Ratio . . . 31
ITU International Telecommunication Union . . . 69
LCIA Low Complexity Intelligibility Assessment . . . 32
LCQA Low Complexity Quality Assessment . . . 29
LSF Line Spectrum Frequency . . . 125
LTASS Long Term Average Speech Spectrum . . . 63
MFCC Mel-Frequency Cepstral Coefficients . . . 125
MIR Music Information Retrieval . . . 138
MNRU Modulated Noise Reference Unit . . . 49
MOS Mean Opinion Score . . . 29
MOS-LQE Mean Opinion Score Estimated with a Parametric Listening Quality algorithm . . . 70
MOS-LQO Mean Opinion Score for Objective Listening Quality . . . 70
MOS-LQS Mean Opinion Score for Subjective Listening Quality . . . 70
NCCF Normalized Cross Correlation Function . . . 46
NICO Non-Intrusive CODEC Identification . . . 29
NISA Non-Intrusive Speech Assessment . . . 29
NISI Non-Intrusive Speech Intelligibility . . . 29
NISQ Non-Intrusive Speech Quality . . . 29
PAMS Perceptual Analysis Measurement System . . . 71
PCA Principal Component Analysis . . . 31
PEFAC Pitch Estimation Filter with Amplitude Compression . . . 47
PESQ Perceptual Evaluation of Speech Quality . . . 29
PLD Power spectrum of Long term Deviation . . . 31
PLP Perceptual Linear Prediction . . . 72
PSQM Perceptual Speech Quality Measure . . . 71
QoS Quality of Service . . . 68
RAPT Robust Algorithm for Pitch Tracking . . . 46
RIR Room Impulse Response . . . 49
RMSE Root Mean Square Error . . . 88
SBS Sequential Backward Selection . . . 38
SCT Speech Corruption Toolkit . . . 30

Mathematical Notation

Symbols and operators

$[.]^{\mathrm{T}}$	Non-conjugate matrix transpose
$\lvert . \rvert$	Absolute value
$*$	Convolution operator
$\mathrm{diag}\{.\}$	Diagonal operator
$e^{(.)}$	Exponential function
$\log(.)$	Logarithm (base 10)
$\max\{.\}$	Maximum function
$\min\{.\}$	Minimum function
$\mu(.)$	Mean function
$\sigma(.)$	Variance function
$\gamma(.)$	Skewness function
$\kappa(.)$	Kurtosis function
f_s	Sampling frequency
ϕ	A single feature for a speech signal
φ	Feature vector for a single speech signal
Φ	Feature matrix
θ_i	Label for the i^{th} signal
Θ	Vector of labels for a set of signals
$\hat{\Theta}$	Vector of estimated labels for a set of signals

Variables and general notation

x	Scalar
$\mathbf{x}$	Vector
$\mathbb{X}$	Matrix
$f(n)$	Function of a discrete variable at time index n
$s(n)$	Clean speech signal
$v(n)$	Noise signal
$y(n)$	Noisy speech signal
$r_p(X,Y)$	Pearson correlation coefficient
$r_s(X,Y)$	Spearman correlation coefficient
F_{cor}	Correlation based feature selection function
F_{PCA}	PCA based feature projection function

Chapter 1

Introduction

SPEECH is the most sophisticated and evolved form of human communication that has received much attention in the field of signal processing resulting in the development of efficient speech communication systems that have become a commodity in the modern age. However when such systems are used in non-ideal conditions they suffer from various degradations, reducing the perceived quality and in more severe cases a loss in the intelligibility of the signal. Such degradations can occur at the conversion medium (microphone) or over the transmission channel (radio link), resulting in additive noise, non-linear effects such as clicks, peak clipping, reverberation and coding artifacts. In this thesis, an ideal listener environment is assumed which is free from background noise or reverberation and all degradations occur in the speaker environment or over the transmission channel (Fig.1.0.1).

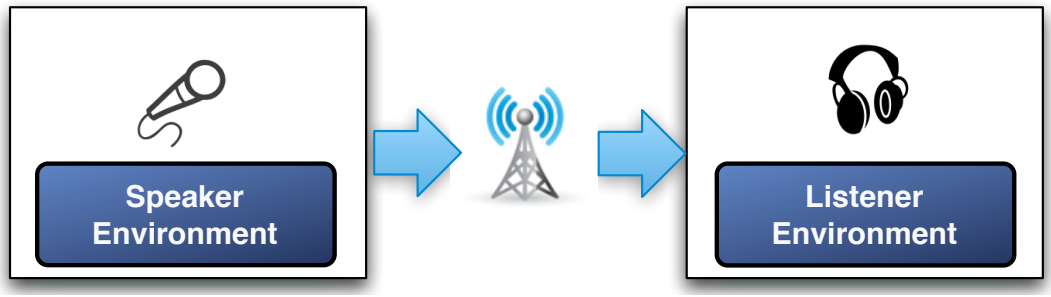

Figure 1.0.1: The listener and speaker environments, separated by a transmission channel.

1.1 Speech assessment and characterization

In this thesis, speech assessment and characterization are defined as the estimation or identification of perceptual or physiological quantities from the degraded speech signal. This includes speech quality and intelligibility assessment and speech Coder-Decoder (CODEC) identification. This task may be performed by human listeners in a subjective experiment or carried out by objective techniques. In general there are two types of objective speech assessment methods (Fig. 1.1.1):

- Intrusive - these methods require the clean speech signal in addition to the degraded speech signal in order to perform the assessment task and are also referred to as double-ended systems.
- Non-intrusive - these methods rely entirely on the degraded speech signal to perform the assessment task.

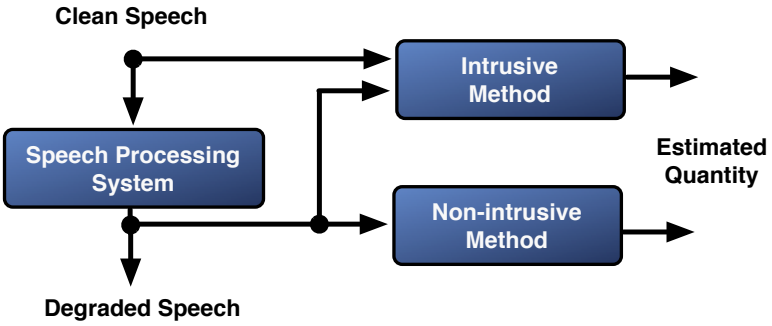

Figure 1.1.1: The two types of objective speech assessment methods: intrusive and non-intrusive. The assessment task typically involves estimation of a perceptual or physiological quantity.

1.2 Context

In the field of law enforcement audio collection, severely degraded audio is commonplace, reducing the intelligence value of the audio by making it unintelligible or inadmissible as evidence in a court of law [110]. In less extreme conditions, a loss in quality is encountered, which has adverse effects on transcription speed resulting in higher operation costs [125]. The areas of poor quality may be processed by a speech enhancement algorithm to improve the quality of

the signal and in such cases it is necessary to ensure that the enhancement does not degrade speech intelligibility (as many speech enhancement techniques have an adverse effect on speech intelligibility [74, 66]). There is a need for non-intrusive objective measures of the signal quality and intelligibility to aid or guide the collection and transcription of law enforcement audio and help optimize the performance of speech communication systems in general. Additionally, identification of the CODEC used in the transmission channel is beneficial for forensic examination of audio, since the presence of a CODEC can effect speaker identification [112, 122] and also help validate the audio collection mechanism.

The problem of speech in noise has received much attention in the literature, however, this development has been largely focused on the telecommunications sector where degradations are typically less severe and the key issue is delivery of a consistent quality of service to the consumer. In the law enforcement scenario, the degradations encountered are typically more severe [125] and many current techniques for speech assessment and characterization are not validated for use in law enforcement scenarios.

1.3 Research Aims

This thesis aims to develop a novel framework for non-intrusive speech assessment in context of law enforcement audio and also investigate the following aspects of a degraded speech signal:

- Degradation mechanism - the degradation processes that result in degraded audio will be investigated and a software tool for simulating the entire speech acquisition, processing and transmission chain will be developed. This tool will be used to generate speech databases with degradations that are relevant in the law enforcement context, which can be used for subjective assessment as well as for validating current and new non-intrusive assessment techniques.
- Signal properties - here the measurable properties of a signal will be investigated for assessment tasks of speech quality, intelligibility and CODEC identification and verification, based on a data-driven machine learning framework.
- Vocabulary - this study will investigate the development of a compact vocabulary for describing the perceptual effects of speech degradations.

1.4 Thesis outcomes and layout

1.4.1 Outcomes

The following is an outline of the research outcomes of this thesis.

The C-Qual database for speech quality assessment in the law enforcement context has been developed and labeled with Mean Opinion Score (MOS) from a subjective experiment comprising of 44 degradation conditions. The results from the subjective experiments and the reliability of the data are presented in Section 2.3.1. The intrusive Perceptual Evaluation of Speech Quality (PESQ) [88] algorithm was validated on the C-Qual database in Section 3.4.4 in order to evaluate the degradation conditions for which PESQ could be used to automatically label large quantities of development and test data. This is a novel approach to non-intrusive algorithm development and has been applied to speech quality and intelligibility assessment. A data-driven, non-intrusive framework for speech assessment is presented in Section 2.4, along with an evaluation of pitch estimation in adverse conditions and a description of the signal features used by the Non-Intrusive Speech Assessment (NISA) framework. The NISA framework has been applied to the following assessment tasks.

1. Speech quality assessment : the problem of per-utterance speech quality is discussed in Section 3.4, where the Low Complexity Quality Assessment (LCQA) [53] algorithm is further developed into the LCQA2 and LCQA-M algorithms, incorporating new features and a two-stage dimensionality reduction scheme. Also, the Non-Intrusive Speech Quality (NISQ) algorithm is proposed in Section 3.4.3 and validated on a large database labeled by the intrusive PESQ algorithm in Section 3.4.7. An initial study on time-varying, non-intrusive speech quality assessment was carried out in Section 3.5 using a concatenation of speech utterances from the C-Qual database.

2. Speech intelligibility assessment : the problem of non-intrusive speech intelligibility assessment is discussed in Chapter 4, where the Non-Intrusive Speech Intelligibility (NISI) algorithm is proposed and tested on a database comprising of additive noise and labeled with the STOI [158] algorithm. The experimental validation of non-intrusive speech intelligibility assessment of noisy speech is presented in Section 4.3.7 and a study on the non-intrusive assessment noise-suppressed speech is presented in Section 4.4.

3. CODEC identification and verification : the Non-Intrusive CODEC Identification (NICO) algorithm is proposed in Section 5.3 and evaluated for the CODEC identification task in Section 5.5.2 and CODEC verification task in Section 5.5.1.

In addition to the modeling of signal properties of the audio, it is important to study the mechanism that produces audio with particular characteristics and the vocabulary that may be used to describe the audible effects of the degradations. These aspects of the signal are presented under the speech description taxonomy in Chapter 6, which is a general framework for characterizing and linking the following aspects of a degraded speech signal:

- Degradation mechanism - a study of the mechanism that produces audio with particular properties. The Speech Corruption Toolkit (SCT) is presented in Section 6.3 as a tool for simulating speech acquisition, processing and transmission system related degradations. The tool allows realistic degradations to be applied to any number of clean speech signals in a repeatable manner.
- Vocabulary - a compact vocabulary for human diagnosis and description of the audio is described in Section 6.4. A large subjective experiment has been conducted with 51 subjects and a 48 label vocabulary was clustered into 10 classes.
- Signal properties - the measurable properties of a signal from which speech assessment and characterization may be performed is described in Section 6.5. The signal properties are encapsulated by the features extracted from the audio signal which may be used as part of machine learning framework to perform a number of tasks, such as speech quality assessment.

1.4.2 Original Contributions

As far as the author is aware, the following are the original contributions of this thesis.

- Development of the C-Qual speech quality database, containing subjective MOS for 44 degradation conditions that were considered important for law enforcement applications.
- Validation of the industry standard PESQ [88] algorithm on the C-Qual database, providing useful insights into the limitations of PESQ in law enforcement scenarios.
- Proposed the use of intrusive algorithms to automatically label development and test data enabling the development and testing of non-intrusive algorithms in scenarios where large quantities of subjectively labeled data is not easily available (such as in law enforcement audio research).
- Development of the data-driven NISA framework for non-intrusive assessment of speech with the addition of novel long-term and short-terms speech features and the use of Classification and Regression Trees (CART) to model the target score. The novel features proposed in the NISA framework include the following
  - Importance weighted Signal to Noise Ratio (iSNR) based on the existing minimum statistics noise estimation algorithm
  - Statistics of the variance and dynamic range of the Hilbert envelope (extracted on a per frame basis)
  - The spectral flatness, spectral centroid and spectral dynamics of the per frame Power spectrum of Long term Deviation (PLD) spectrum.
  - The long term deviation features based on the mean PLD spectrum for the entire audio.
- Development of the NISQ algorithm for speech quality assessment and the NISI algorithm for speech intelligibility assessment based on the NISA framework with regression trees and shown to perform favorably on the databases tested.
- Extension of the existing LCQA algorithm with the addition of novel speech features and a two step dimensionality reduction scheme using Principal Component Analysis (PCA) and feature correlations, resulting in the LCQA2 and LCQA-M algorithms, which outperform the baseline LCQA and ITU standard P.563 methods in non-intrusive speech quality assessment.

- Investigation into non-intrusive assessment of time-varying speech quality with block based extensions to the LCQA, LCQA2, LCQA-M and NISQ algorithms. Block sizes in the range 0.5 to 8.0 seconds were investigated and new results on time-varying speech quality have been obtained.
- Extension of data-driven algorithms for non-intrusive assessment of speech intelligibility, including an investigation into non-intrusive assessment of the effects of spectral subtraction on speech intelligibility. The Low Complexity Intelligibility Assessment (LCIA) algorithm was published and represents a first attempt at non-intrusive intelligibility assessment (as far as the author is aware).
- Development of NICO algorithm for non-intrusive CODEC identification and verification from the NISA framework using classification trees. Current CODEC identification and verification techniques have focused on clean speech conditions, this research extends the problem by considering additive noise conditions. The proposed NICO algorithm is shown to be robust to additive noise effects (based on the noises tested).
- Development of the speech description taxonomy framework for characterizing and linking various aspects of a degraded speech signal, including the degradation mechanism, the measurable signal properties and a compact vocabulary for human diagnosis and description of the audio.
- Development of a tool (Speech Corruption Toolkit) for simulating speech acquisition, processing and transmission system related degradations.
- Development of a 10 class speech degradation vocabulary using subjective test data from a large experiment and hierarchical clustering of the initial 48 word vocabulary.

1.4.3 Publications

The following is a list of publications related to the research presented in this thesis:

1. D. Sharma and P. A. Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment," in Proc. European Signal Processing Conf. (EUSIPCO), Glasgow, Aug. 2009.
2. D. Sharma, G. Hilkhuysen, N. D. Gaubitch, M. Brookes, and P. A. Naylor, "C-Qual - a validation of PESQ using degradations encountered in forensic and law enforcement audio," in Proc. AES Conf. on Audio Forensics, Hillerød, Denmark, Jun. 2010.

3. D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in Proc. European Signal Processing Conf. (EUSIPCO), Denmark, Aug. 2010.

4. D. Sharma, P. A. Naylor, N. Gaubitch, and M. Brookes, "Short-time objective assessment of speech quality," in Proc. European Signal Processing Conf. (EUSIPCO), Barcelona, Aug. 2011.

5. D. Sharma, P. A. Naylor, N. D. Gaubitch, and M. Brookes, "Non intrusive CODEC detection algorithm," in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Mar. 2012.

6. D. Sharma, G. Hilkhuysen, P. A. Naylor, N. D. Gaubitch, M. Huckvale and M. Brookes, "Descriptive Vocabulary Development for Degraded Speech", to appear in Proc. INTERSPEECH, Portland, Sep. 2012.

The following is a list of publications that have also been produced over the past 4 years, but not presented in the thesis:

- N. D. Gaubitch, M. Brookes, P. A. Naylor, and D. Sharma, "Bayesian adaptive method for estimating speech intelligibility in noise," in Proc. AES Conf. on Audio Forensics, Hillerød, Denmark, Jun. 2010.

- P. A. Naylor, N. D. Gaubitch, D. Sharma, G. Hilkhuysen, M. Huckvale, and M. Brookes, "Intelligibility estimation in law enforcement speech processing," in Proc ITG Conf on Speech Communication, Bochum, Germany, Oct. 2010.

- N. Gaubitch, M. Brookes, P. A. Naylor, and D. Sharma, "Single-microphone blind channel identification in speech using spectrum classification," in Proc. European Signal Processing Conf. (EUSIPCO), Barcelona, Aug. 2011.

1.4.4 Layout

The remainder of the thesis is organized as follows. In Chapter 2 an overview of machine learning is presented along with the data-driven NISA framework and the C-Qual database. This is followed by the application of the NISA framework for speech quality assessment in Chapter 3, for speech intelligibility assessment in Chapter 4 and for the task of CODEC identification and verification in Chapter 5. In Chapter 6, the speech description taxonomy is presented and finally the conclusions and future work in Chapter 7. The thesis structure is illustrated in Fig. 1.4.1 with the thesis chapters described in relation to the taxonomy framework outlined in the previous section.

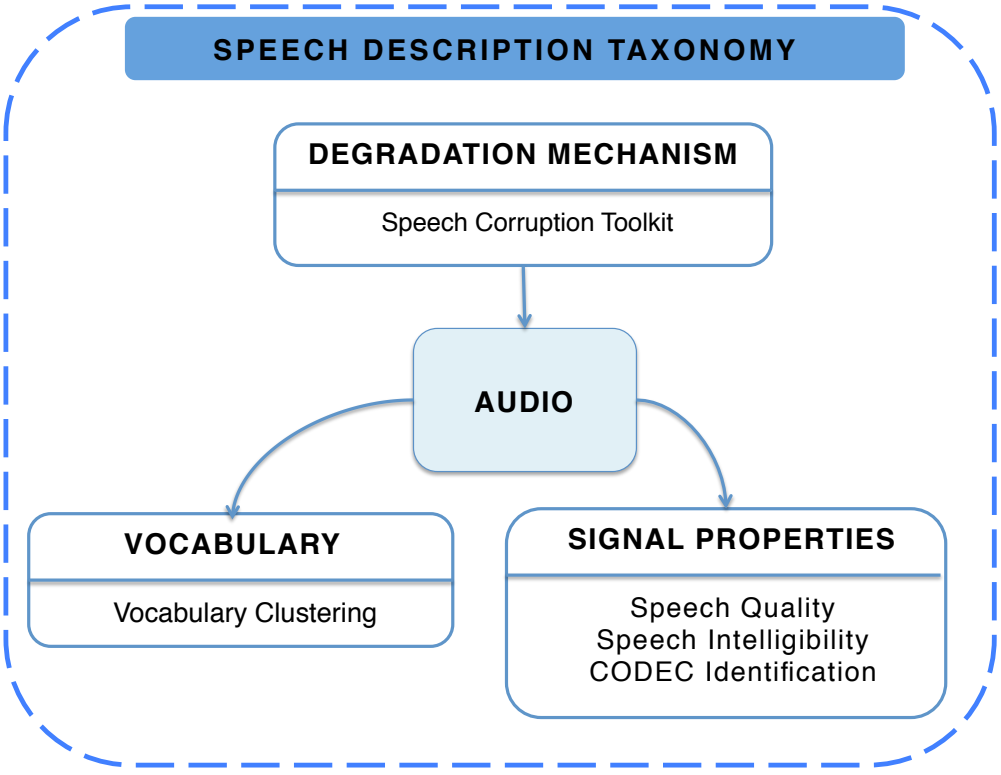

Figure 1.4.1: The three aspects of an audio signal as part of the speech description taxonomy (SDT): Degradation mechanism, Vocabulary and Signal properties. The signal properties are concerned with the assessment task of speech quality, intelligibility and CODEC identification, described in Chapters 3, 4 and 5 respectively. The SDT framework is outlined in Chapter 6.

Chapter 2

Non-Intrusive Speech Assessment Framework

THIS chapter presents a review of machine learning and speech production and analysis background necessary for development of the non-intrusive speech assessment framework in Sections 2.1 and 2.2 respectively. A requirement for developing and validating data-driven machine learning algorithms is to have labelled data with similar degradations as those expected in the problem domain. In Section. 2.3, labelled datasets are presented, including the C-Qual speech quality database for law enforcement audio. Then, having established the basis for a data-driven approach and appropriate databases, the NISA framework is presented in Section. 2.4. The work in this chapter relates in part to the following publications [147, 146].

2.1 Machine learning background

A typical machine learning approach to non-intrusive assessment is illustrated in Fig. 2.1.1. The first stage is a pre-processing of the input signal which in the context of speech processing may involve energy normalization, voice activity detection and segmentation of the signal into short time frames. The second stage is a feature extraction that aims to extract robust features from the signal that are invariant to certain transformations (such as gain manipulation) and that can capture those characteristics of the signal that are of interest [31]. This is followed by a dimensionality reduction stage, where a subset of features are selected or a linear combination of the current features are combined into a smaller feature set. This may be done before

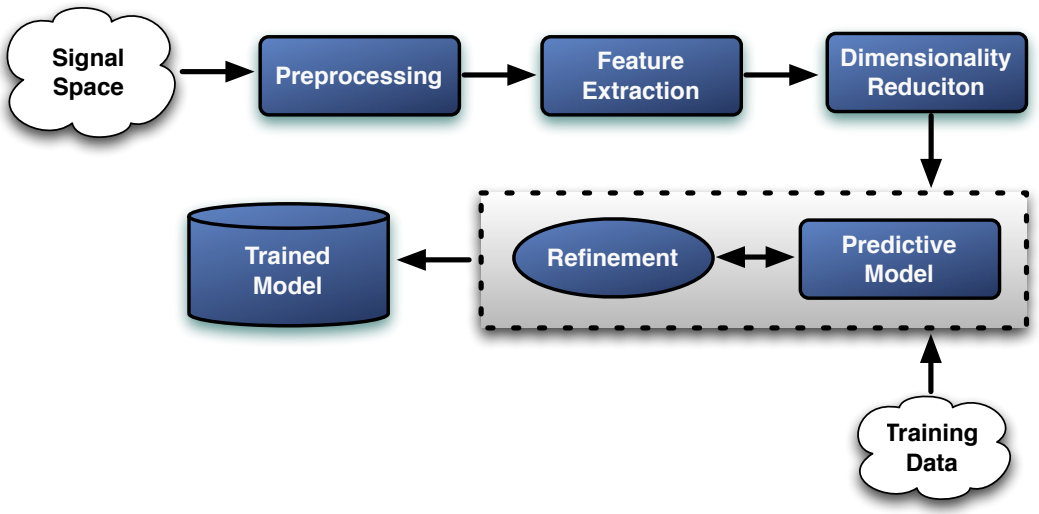

Figure 2.1.1: An example machine learning framework used in non-intrusive assessment.

model training or may be an inherent part of the model training process. The final step is model training, where a model is trained on some prior data, known as training data. Due to the dependency of a machine learning algorithm on the quantity and quality of training data [92], such techniques are also known as data-driven techniques. The training may be supervised, where the ground truth label is available for each feature vector and the training algorithm can compute the cost of misclassification explicitly. However, in some situations the ground truth corresponding to each feature vector is not known (unsupervised training) and the training algorithm must decide on the groupings for each feature vector by applying a clustering approach. The performance of a machine learning algorithm is affected by a combination of learning data quality and quantity, the number and quality of the features and model complexity [92].

The features derived from the signal are referred here as explanatory variables and the ground truth labels are the response variables. The machine learning task is the modeling of the relationship between the explanatory and response variables. When the response variable is categorical, a classification model is constructed and when it is a real valued number, a regression model can be constructed. Let the definition of a training set, L, be as follows

$$L = [\Phi, \Theta] = [(\varphi_1, \theta_1), ..., (\varphi_L, \theta_L)],$$

where Φ is the feature matrix for the training data, defined as

$$\Phi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{N_L} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \phi_1(1) & ,..., & \phi_{N_L}(1) \\ \phi_1(2) & ,..., & \phi_{N_L}(2) \\ \vdots & \vdots & \vdots \\ \phi_1(N_f) & ,..., & \phi_{N_L}(N_f) \end{bmatrix}, \tag{2.1.1}$$

and φ_i is the feature vector for the i^{th} signal, defined as

$$\varphi_i = [\phi_i(1), \phi_i(2), .., \phi_i(N_f)]^{\mathrm{T}}, \tag{2.1.2}$$

and Θ (either a real or categorical response variable) denotes the corresponding labels for N_L signals. The notation, Φ_i denotes a vector containing all examples of the i^{th} feature (i.e. the i^{th} row of Φ). The feature matrix, Φ has dimensions $N_f \times N_L$, where N_f is the number of features and N_L is the number of signals in the training set.

2.1.1 Dimensionality reduction

An important issue in machine learning is that of dimensionality reduction. This is carried out to reduce the number of features (N_e from N_f) so that there are fewer parameters to train, help improve generalization performance and reducing the computational complexity of the machine learning system [18, 31, 92]. The two categories of dimensionality reduction relating to the signal features are described in the following subsections.

2.1.1.1 Feature subset selection

The aim of feature subset selection is to select a lower dimensionality subset of features from the original feature vector [91]. An exhaustive search through all possible subsets guarantees an optimal feature subset, however this is computationally infeasible to perform in practice [128]. A simple technique is to evaluate the correlations between the features and the target labels and select those features that have a high correlation with the target labels but are poorly correlated with other features [57]. The following is the definition of the correlation based feature selection method for feature i

$$F_{cor}(i) = \frac{r_p(\Phi_i, \Theta)}{\sum_{j \neq i} r_p(\Phi_i, \Phi_j)}, \tag{2.1.3}$$

where $r_p(\Phi_i, \Theta)$ is the Pearson correlation coefficient between the vector containing all examples of the i^{th} feature and the vector of labels for each signal (Θ), defined as

$$r_p(X,Y) = \frac{\sum_{n=1}^{N}(X_n - \mu_X)(Y_n - \mu_Y)}{\sum_{n=1}^{N}(X_n - \mu_X)^2 \sum_{n=1}^{N}(Y_n - \mu_Y)^2}, \tag{2.1.4}$$

where μ_X and μ_Y are the means of the vector quantities X and Y respectively, and N is the dimension of X and Y. The Pearson correlation coefficient measures the linear relationship between two variables.

The sequential search methods evaluate various subsets of the original features and rejects those features that have a small effect on the machine learning performance. These techniques provide a tradeoff between computational complexity and optimum feature subset selection [92]. The Sequential Forward Selection (SFS) algorithm [129] begins with an empty set and evaluates each feature in turn, adding the best feature at each iteration, where the best feature is the one that improves system performance when combined with features already in the subset of features. The counterpart to this method is the Sequential Backward Selection (SBS) algorithm [129], which begins with a full set of features and successively prunes away the least significant features. These methods are sub-optimal as once a feature is removed or added to the final subset, it is not re-evaluated.

The floating search methods proposed by Pudil *et al.* [129] are extensions to the basic sequential search methods, where a number of forward steps are applied after each backward step (Sequential Floating Backward Selection (SFBS)) and the number of forward or backward steps is controlled dynamically through a threshold on the improvement in system performance. The counterpart to the SFBS method is the Sequential Floating Forward Selection (SFFS), which begins with an empty feature set and sequentially applies a number of forward and backward steps, as with SFBS.

2.1.1.2 Feature Projection

In this approach, the original features are transformed into a lower dimensional feature space by a linear combination of the original features. The PCA (also known as the Karhunen Loeve transform) is the best known linear feature extraction algorithm [92] that re-expresses the feature space as an orthogonal basis by linear combination of the original features [150]. It is a non-parametric, unsupervised method as it evaluates the feature vectors without knowledge of their labels [18]. An eigenvalue decomposition of the feature covariance matrix is performed and the eigenvectors corresponding to N_e eigenvalues are used perform the dimensionality reduction.

Let the covariance matrix of the feature matrix, Φ, be defined as

$$\underline{\mathbb{C}} = \frac{1}{N_L}\Phi\Phi^{\mathrm{T}},$$

where $\underline{\mathbb{C}}$ is an $N_f \times N_f$ symmetric covariance matrix, then the eigenvalue decomposition of $\underline{\mathbb{C}}$ can be defined as

$$\Lambda = V\underline{\mathbb{C}}V^{\mathrm{T}},$$

where Λ is a diagonal matrix of eigenvalues and V is an $N_f \times N_f$ matrix of the eigenvectors for the feature covariance matrix $\underline{\mathbb{C}}$. The PCA based feature extraction measure is given by

$$F_{PCA} = V_e^{\mathrm{T}} \cdot \Phi, \tag{2.1.5}$$

where V_e^{T} is an $N_e \times N_f$ matrix of the N_e eigenvectors and Φ is the feature matrix.

2.1.2 Classification and Regression Trees

CART [20] is a recursive partitioning algorithm with a number of desirable properties and has been applied to a number of problems, including customer credit scoring [97], ecological data analysis [29] and speech quality assessment [180]. The partitioning results in a decision tree that can be applied to any data structure (discrete or continuous) and handles dimensionality reduction automatically as part of the tree construction process. The final CART model has a low run-time complexity and a human readable format. There are three main elements to a CART model construction:

1. Node splitting rule : a criterion that decides the binary splits of Φ into child nodes given a training set.
2. Stopping criterion : a rule to decide when to stop growing a tree.
3. Class assignment : a rule for assigning a label to a terminal node

The CART method begins by growing an oversize tree (one that is suboptimal) and then applies a pruning operation to optimize the tree, typically using cross-validation [106]. The motivation for this approach is that it is difficult to agree on a stopping criterion in the tree growing process

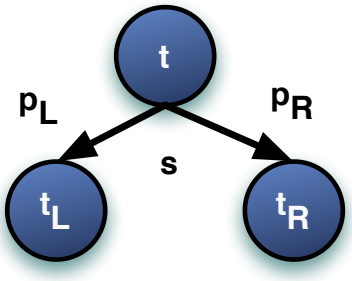

Figure 2.1.2: A simple tree structure with a parent node t split into two child nodes t_R and t_L using splitting criterion s.

that is guaranteed to be optimal, as although splitting a node may have a small change in the node splitting criterion, there may be a split further down that branch that gives a significant reduction. At each node the algorithm searches exhaustively through each of the features for the optimal split and the feature with the best split is selected. These concepts are further explained in Section. 2.1.2.1 for classification trees (where the response variable is categorical) and regression trees in Section. 2.1.2.2 (where the response variable is continuous).

2.1.2.1 Classification trees

Let the N_f dimensional feature vector for a single audio signal be denoted as φ_i, defined as a column in the feature matrix Φ (Equation 2.1.2) and let the vector of classes be denoted as Θ, then the general classification problem is a function $d(\varphi) = \hat{\theta}$ that maps every column in Φ to an element in Θ [20]. Figure. 2.1.2 shows a simple decision tree with a parent node t split using criterion s into a right leaf node t_R and a left leaf node t_L, with p_L and p_R being the proportion of the data at t that go the the left and right leaf nodes respectively.

Given a classifier $d(\varphi)$, let $R^\star(d)$ denote the true misclassification rate of $d(\varphi)$. The most accurate estimate for $R^\star(d)$ is obtained with an infinitely large set of labeled data that could be considered to be the population from which the training data is sampled. As this is typically not feasible, three estimates of $R^\star(d)$ are defined as follows. Let $R_e(d)$ denote the re-substitution estimate, defined as

$$R_e(d) = \frac{1}{N_L} \sum_{n=1}^{N_L} \Upsilon(d(\varphi_n), \theta_n), \tag{2.1.6}$$

where N_L is the number of audio signals in L, θ_n is the class label for the n^{th} signal and Υ is an

index function defined as

$$\Upsilon(a,b) = \begin{cases} 1 & \text{if } \neq b \\ 0 & \text{otherwise} \end{cases} .$$

As the re-substitution metric is computed using the entire training data, it is typically an underestimate of the true misclassification rate. When a large amount of training data is available, the training data may be split into a training and a test set ($L = [L_1 \cup L_2]$) then the test sample estimate of the misclassification error is defined as

$$R_{ts}(d) = \frac{1}{N_{L2}} \sum_{(\varphi_n,\theta_n)\in L_2}^{N_{L_2}} \Upsilon(d(\varphi_n),\theta_n). \tag{2.1.7}$$

Another metric often used is the V-fold cross validation metric where the training set is divided into V subsets ($L = [L_1 \cup L_2 ... \cup L_V]$) defined as

$$R_{cv}(d) = \frac{1}{V} \sum_{v=1}^{V} \left(\frac{1}{N_V} \sum_{(\varphi_n,\theta_n)\in L_v} \Upsilon(d^v(\varphi_n),\theta_n) \right), \tag{2.1.8}$$

where $N_V = N_L/V$ and $d_v(\varphi_n)$ is a classifier trained on $N - N_V$ data. A value of 10 for V has been heuristically determined to give satisfactory estimates of the true misclassification rate [20].

The splitting of a node is dictated by a reduction in the impurity of the data at the node, which is a measure of the diversity of target classes for the data at the node. Following this definition, a node is defined to be pure if all the data in that node belong to the same class. Let $i(t)$ be the node impurity function, then the change in impurity due to splitting node t using criterion s can be decomposed as

$$\Delta i(s,t) = i(t) - (p_L \cdot i(t_L) + p_R \cdot i(t_R)).$$

Let S be the set of all candidate splits at a node and let $S^\star$ be the optimal split, i.e. the one that gives the largest decrease in impurity at node t. Then the change in impurity at t due to $S^\star$ is

$$\Delta i(s^\star,t) = \max_{s\in S} \Delta i(s,t).$$

The Gini diversity index [20] is a splitting criterion often used in classification trees, defined as

$$i(t) = 1 - \sum_j p^2(\theta = j|t),$$

where $p(\theta = j|t)$ is the proportion of data belonging to class $\theta = j$ at node t. A node with all data belonging to the same class is referred to as a pure node and has a Gini index equal to 0. The initial tree is grown until the current node is pure or if there are fewer than 10 observations in the node (heuristically determined) [20].

The second step is the pruning of the initial tree $\mathbb{T}$ into the optimal tree $\mathbb{T}_{Opt}$ by merging leaf nodes at the same level in the tree. A 10-fold cross validation of the training set is used for pruning the initial classification tree [20], where the cost of pruning $\mathbb{T}$ at different levels is calculated for each of the 10 partitions and the pruning level that produces the smallest tree within 1 standard error of the minimum cost subtree is selected. The final classification tree is obtained by pruning $\mathbb{T}$ at this heuristically determined level.

2.1.2.2 Regression trees

The process of growing and pruning a regression tree follows similar concepts as the classification tree procedure. The response variables (Θ) in this case are real valued numbers. The objective in regression is to make a real valued function that predicts the values Θ given a feature vector φ. This function is denoted as $d_r(\varphi)$. The feature space is successively partitioned into a number of nodes and at each terminal node the predicted value ($\hat{\theta}$) is a constant [20]. As with the classification tree algorithm, an important question is the estimation of the true prediction error $R^\star(d)$ for a predictor $d_r(\varphi)$. In the context of tree based regression, the typical metric for this is the mean square error

$$R^\star(d) = E(\Theta - \hat{\Theta}))^2, \tag{2.1.9}$$

where Θ is the vector of ground truth values and $\hat{\Theta}$ is the vector of estimated values. The re-substitution estimate of $R^\star(d)$ is given by

$$R_e(d) = \frac{1}{N_L} \sum_{n=1}^{N_L} (\theta_n - d_r(\varphi_n))^2, \tag{2.1.10}$$

where N_L is the number of signals in L. The value of $\hat{\theta}$ that minimizes $R_e(d)$ is the mean of the values θ_n at node t

$$\bar{\theta}_t = \frac{1}{N_t} \sum_{n \in t} \theta_n \tag{2.1.11}$$

where N_t is the total number of data samples at node t. Let the regression impurity function for node t be defined as

$$R(t) = \frac{1}{N_t} \sum_{n \in t} (\theta_n - \bar{\theta}_t)^2.$$

Let $\Delta R(s,t)$ be the reduction in $R(t)$ when node t is split into t_L and t_R nodes

$$\Delta R(s,t) = R(t) - (R(t_L) + R(t_R)),$$

and $R(t) \geq R(t_L) + R(t_R)$. Let $s^\star$ be the optimal split, defined as the split in S (the set of all possible splits) that causes the largest decrease in $R(t)$, then

$$\Delta R(s^\star, t) = \max_{s \in S} \Delta R(s,t).$$

The initial regression tree is grown until all nodes are pure (all values at the node are the same) or if there are fewer than 10 observations in the node (heuristically determined). The final value at a terminal node is the average of the labels at that node (Equation. 2.1.11). The pruning of the initial regression tree follows the classification tree algorithm, that is, a 10 fold cross validation is used to test the performance of the tree and the cost of pruning $\mathbb{T}$ at different levels is calculated for each of the 10 partitions. The optimal pruning level is the tree that achieves errors within 1 standard error of the minimum cost sub-tree [20] and the final regression tree is obtained by successively pruning $\mathbb{T}$ at this heuristically determined level.

2.1.3 Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is a parametric model for continuous variables and has been applied to a number of speech processing applications such as speaker identification and verification [135] and speech quality estimation [53]. A GMM is a linear combination of M Gaussian densities of the form

$$p(\varphi | w, \mu, \Sigma) = \sum_{m=1}^{M} w_m \times \mathcal{N}(\varphi | \mu^{(m)}, \Sigma^{(m)}),$$

where φ is a feature vector, $\mathcal{N}(\varphi|\mu^{(m)},\Sigma^{(m)})$ is a multivariate Gaussian density and w is the mixture coefficient vector with the following property

$$\sum_{m=1}^{M} w_m = 1.$$

An M component GMM is fully specified by the mixture weights (w), means (μ) and covariance matrices (Σ). A maximum likelihood estimate of the parameters can be used to learn the GMM parameters using the iterative, expectation-maximization (EM) algorithm [30].

2.2 Speech production and analysis background

This section presents an overview of speech production in Section 2.2.1 followed by an overview of linear prediction in Section 2.2.2 and pitch estimation in Section 2.2.3.

2.2.1 Speech production

The key aspects of the physiology of the human speech production system are depicted in Fig. 2.2.1. The vocal tract begins at the glottis and ends at the lips and the nasal tract begins at the velum and ends at the nostrils [133]. The tongue is used to alter the vocal tract shape and the velum controls air flow through the nasal cavity. The air flow from the lungs into the glottis cause the vocal folds to vibrate during voiced sounds, producing a quasi-periodic excitation for the vocal tract. The frequency of vibration of the vocal folds depends on the tension in the vocal folds and the pressure from the lungs [124] and are responsible for the auditory sensation of pitch. The resulting pressure waveform is shaped by the frequency selectivity of the vocal tract and the resonant frequencies of the vocal tract are referred to as formants [133]. Additionally, some speech sounds can also be classified as unvoiced, which arise when air is forced through a constriction in the vocal tract (usually towards the mouth) at a high velocity, resulting in turbulent air flow to excite the vocal tract [133]. The time-varying speech waveform is thus articulated by the movements of the tongue, jaw, lips, and the velum.

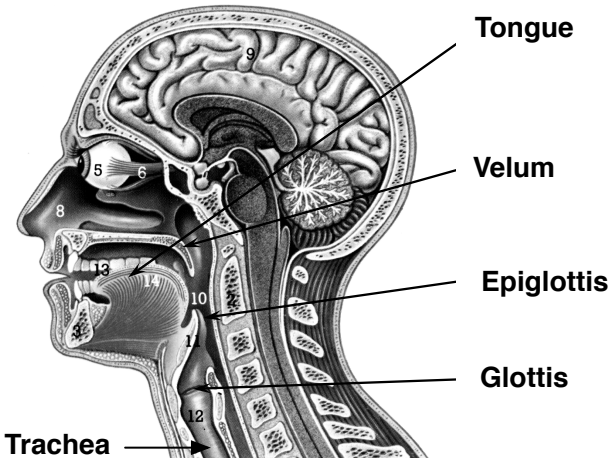

Figure 2.2.1: A simplified diagram of the human speech production system [124]. The human speech production system is driven by the air flow from the lungs into the vocal tract, which is the section of the tube from the glottis to the lips.

2.2.2 Linear prediction

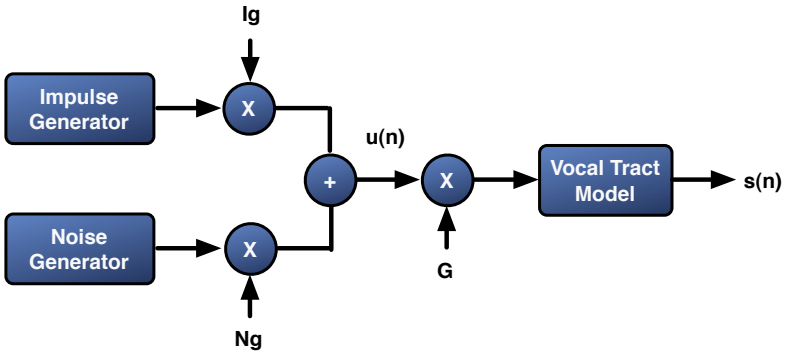

Figure 2.2.2: A block diagram of the source-filter model of speech production. The excitation signal is shown as a mix of noise with a noise gain, Ng (representing unvoiced speech) and an impulse train with a separate gain, Ig (representing voiced speech).

Linear predictive analysis is based on the source-filter model [43] for speech synthesis (Fig. 2.2.2), where a speech signal $s(n)$ can be modeled as the sum of a linear combination of previous speech samples (an all-pole vocal tract model) and an excitation signal $G \cdot u(n)$ as follows

$$s(n) = \sum_{j=1}^{p} \alpha_j s(n-j) + G \cdot u(n).$$

The prediction error for an estimated vocal tract filter is defined as [10]

$$e(n) = s(n) - \sum_{j=1}^{p} \alpha_j \cdot s(n-j),$$

where $e(n)$ is the prediction error and α_j are the estimated prediction coefficients. The prediction coefficients are estimated by the minimizing the mean square of the prediction error, leading to the following set of equations [133]

$$\sum_{j=1}^{p} \alpha_j \cdot \psi_{i,j} = \psi_{i,0}, \tag{2.2.1}$$

$$\psi_{i,j} = \sum_{n} s(n-i)s(n-j)$$

where $i = 1, ..., p$ and the set of p equations with p unknowns (2.2.1) may be solved using for example the autocorrelation or covariance methods [132].

2.2.3 Pitch estimation

The pitch of a speech signal has been used as a feature in a number of speech processing applications, including speech coding [99] and speech quality assessment [80, 53]. The perceived pitch of a speech signal is an inherently subjective quantity which correlates well with the fundamental frequency of the signal [159]. Pitch is associated with periodic excitation that arises due to the oscillation of the vocal folds which modulates the airflow through the glottis. This modulation of the airflow serves as excitation for the vocal tract during voiced speech. The estimation of the pitch of a speech signal from the speech waveform alone is a challenging problem due to the quasi-periodic nature of pitched speech and mixed nature of the excitation [134].

The Robust Algorithm for Pitch Tracking (RAPT) [159] is a frame based algorithm which uses Normalized Cross Correlation Function (NCCF) as the primary candidate generation function and uses dynamic programming to refine the pitch estimation. The NCCF is the autocorrelation function normalized by the energy of the input signal and is the most computationally

expensive operation in RAPT and so the algorithm performs the NCCF in a two pass process. A down-sampled version of the input signal is used to estimate the first set of candidate peaks, followed by a high resolution (full sample rate) NCCF around the candidates of interest. The YIN [28] algorithm uses an autocorrelation based difference function as the candidate generator in conjunction with a number of optimization steps. While the autocorrelation based methods aim to maximize the product between the waveform and its delayed duplicate, the difference function based approach aims to minimize the difference between the waveform and its delayed duplicate. The YIN algorithm was shown to have a good performance in a number of degradations in [147], including additive noise and reverberation.

The LCQA algorithm [53] utilizes an autocorrelation based pitch feature [99], which is computed by searching for a peak in the autocorrelation function around the pitch period range of interest. This method has a low complexity but also has a poor performance in adverse conditions[1].

The Pitch Estimation Filter with Amplitude Compression (PEFAC) algorithm [52] is a robust method for pitch detection in adverse conditions. The method operates in the frequency domain by transforming the signal using the Fast Fourier Transform (FFT) and uses a comb filter to find the possible pitch estimates for the frame. An optimization process then selects the best pitch estimate from the set of possible pitch candidates using dynamic programming. This method is used in the NISA framework as it gives consistent performance at high levels of additive noise, as shown in Section 2.4.2.1.

Ground truth estimation

In order to evaluate pitch estimation techniques for law enforcement scenarios, a ground-truth pitch label is required. This can be accomplished with the use of a speech database that contains contemporaneous recordings of Electroglottograph (EGG) signals for spoken speech material[2], from which Glottal Closure Instants (GCI) can be extracted and then mapped to the pitch [64, 161]. The SIGMA [160, 161, 21] algorithm operates on an EGG signal and identifies the GCI and Glottal Opening Instants (GOI) for voiced speech. The SIGMA algorithm is based on a stationary wavelet transform preprocessor, with a group delay function as the peak detection function. GMMs are used to classify true and false detections to further improve the performance of the algorithm. The SIGMA algorithm has been shown to provide an average GCI hit rate greater than 99% when compared to hand-labeled GCIs.

[1] See Section 2.4.2.1 for comparative results.
[2] The SAM database [25] has such material.

2.3 Labelled datasets for Law Enforcement Degradations

A key component of a data-driven algorithm is the requirement of labelled training data with the characteristics of the problem domain. Additionally, labelled datasets are required for validating current algorithms in the context of law enforcement as many current systems are only validated for the telecommunications industry. In Section 2.3.1 the C-Qual database [146] is described, which contains the types and levels of degradations commonly found in law enforcement audio and has been subjectively scored with speech quality ratings (MOS) with high data reliability. However, since human based subjective testing is time consuming and expensive [139], an automatic method for generating labelled data is presented in Section 2.3.2.

2.3.1 C-Qual database

The C-Qual database [146] is a subjectively labelled database comprising of 44 degradation conditions, representative of the types and levels of degradations encountered in law enforcement audio. The database has been labelled with speech quality ratings according to the ITU-T P.800 protocol [75]. The remainder of this section outlines the design, results and reliability of the database.

Subjects

Subjective quality scores were obtained from 24 subjects (same as the number of subjects in P.23 [77] experiments), who were native speakers of British English. The criteria for a native speaker was that the subject completed their education (including primary education) in an English medium school in the UK. In addition, the listener selection process required subjects to have a non-technical background to ensure them to be naïve to the effects of the degradations presented[3]. All subjects were verified to be normal hearing, defined here as having hearing thresholds of 20 dB HL or below at octave frequencies ranging from 125 to 8000 Hz. A hearing threshold of 20 dB HL is typically considered to be normal [3]. The subjects were paid for their participation in the experiments.

[3]None of the subjects had taken part in a study of speech quality assessment before.

Stimuli

The English subset of the ITU-T P.23 [77] database was used for the speech material, consisting of a pair of utterances separated by a small pause. The average duration of each stimuli was approximately 10 seconds. Speech from two male and two female speakers were included in the stimuli. The speech level was adjusted to give all files the same level using the ITU-T P.56 [82] method, before further processing and resampled to 16 kHz. Six types of degradations were included and the level of distortion for each degradation type was chosen to cover the range of MOS from 1 to around 5 using the PESQ algorithm. The following is a description of the degradations in the database

1. Additive noise: car, babble and hum noise were included at seven Signal-To-Noise Ratio (SNR). Car and babble noise were added to the clean speech at -16, -8, 0, 8, 16, 24 and 32 dB SNR and hum noise was added at -30, -20, -20, 0, 10, 20 and 30 dB SNR.

2. Reverberation: Room Impulse Response (RIR) from two rooms were included with different microphone to source distances. The Multichannel Acoustic Reverberation Database at York (MARDY) room [175] was included with 3 microphone-to-speaker distances and the Imperial room with 2 distances. The reverberation time (T_{60}) for the MARDY and Imperial rooms were calculated [142] to be 1.35 and 1.02 seconds respectively.

3. Coloration: three types of shelf-filters were included: low cut (Fig. 2.3.1), high boost (Fig. 2.3.1) and anti-clockwise spectral tilt (Fig. 2.3.2).

4. Peak clipping: symmetric hard clipping was applied at four levels (-8, -12, -16 and -20 dBFS).

5. Clicks: five levels of temporal erasures were included. These were generated by applying a short rectangular window of zeros to the speech signal. The number of clicks with a duration of 20 ms was 2, 7 and 16. In addition 150 and 440 clicks were added with a shorter duration of 5 ms. This resulted in click durations of 40, 140, 320, 750 and 2200 ms respectively. The position of clicks was randomly determined and confined to the speech active regions only (i.e. no clicks were added in the pause segments).

6. Modulated Noise Reference Unit (MNRU) [76]: six levels of amplitude modulation are applied to the speech. These were included to compare the results from this study with results obtained in the P.23 database[4].

[4] English partition of Experiment 1 (results from the BNR laboratory).

The 44 degradation conditions were applied to speech from 4 speakers to create 5 blocks of stimuli, resulting in a total of 220 audio stimuli. The number of test conditions was guided by the constraint of a 1 hour test duration for each subject. Further details of the degradation conditions are presented in Appendix A.

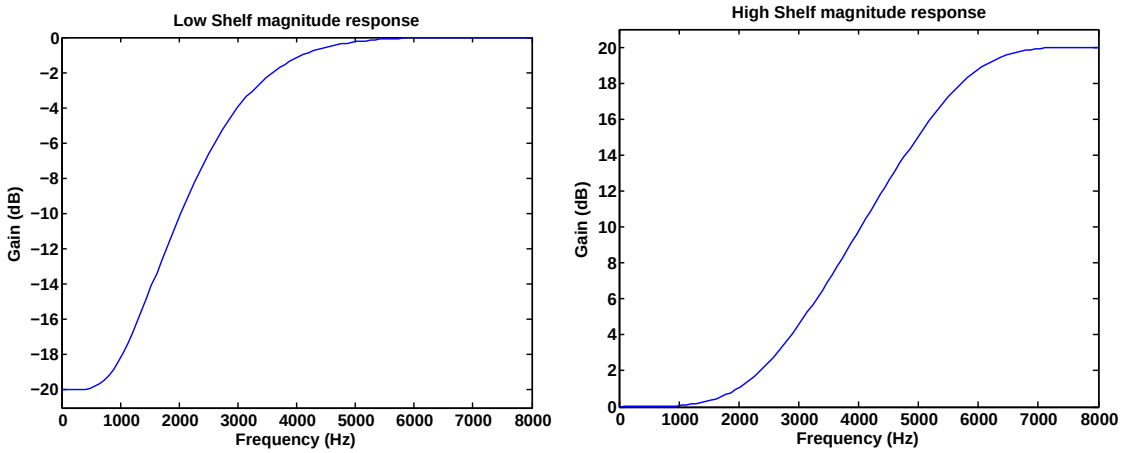

Figure 2.3.1: The low cut (left) and high boost (right) shelf filter magnitude responses.

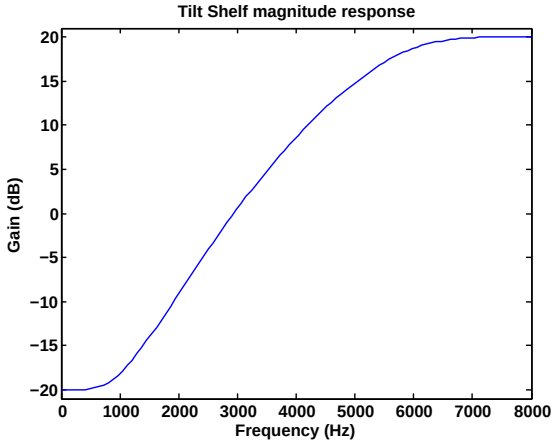

Figure 2.3.2: The spectral tilt filter's magnitude response using a low-cut and a high-boost shelf filter, covering a 40 dB magnitude range.

Task

The task for the subjects was to listen to the stimuli and give a MOS on a scale from 1 to 5, based on the ITU-T P.800 [75] protocol. The subjects were asked to score the overall effect on

a single dimension[5]. The stimuli were presented in five sessions to each listener, each session containing 44 audio examples and the first one being a practice session designed to help the subjects get familiarized with the experiment. The subjects were given a small introduction on the context of the research and were instructed to listen to the audio and give a response on the 5 point MOS scale for the overall quality of the sound. The written instructions are presented in Fig. 2.3.3. A randomization of stimuli presentation order within sessions and between subjects was applied. All stimuli were presented at 60 dB SPL (fixed speech level) over Sennhieser HD 650 headphones and listening tests were conducted in a sound-proof booth.

C-Qual Database

In this experiment you will be listening to short sentences through the headphones provided and giving your opinion on the quality of audio. On the PC user interface there is a button to play the audio (you may replay the audio if you need more time to decide the score) and give your response on the 5 point opinion scale provided. Please take a moment to look at the quality score descriptions on the screen and feel free to ask any questions you have about the scale.

To begin with, the first session will be a practice session to allow you to get an idea of the types and levels of degradations that will be played through the experiment. After the practice session, you may take a small break and ask any further questions you may have. We will give you more details on the background and motivation for this study after the experiment is completed, this is to allow the response to come from naieve listeners. Thanks for taking part in this study!

Figure 2.3.3: Instructions for the C-Qual database.

Results

This section presents the reliability of the subjective scores obtained by the listening tests and the effect of the degradations on the MOS. The results from the listening tests highlight the relationship between the perceived quality of speech for each degradation condition (level and type of degradation). The box-plots present the median (central line in the box), the 25th and 75th percentiles are represented as the limits of the box, the dashed lines present the extreme data points and outliers are plotted as a '+'. The results for the subjective MOS are given as the average across all 24 subjects and 4 speakers and referred to as condition averaged MOS.

Figure. 2.3.5 shows the relationship observed for hum noise with SNRs ranging from -30 to 30 dB. The overall relationship for car and babble noise resembles a 'sigmoid' curve observed

[5]Absolute Category Rating (ACR)

in the context of intelligibility testing, where there is a linear mid region and a saturation effect at the low and high ends (Fig. 2.3.4). A similar relationship is observed between different levels of peak clipping and clicks with subjective MOS (Fig. 2.3.6). The effect of reverberation for the MARDY database is shown in Fig. 2.3.8 and for the Imperial room in Fig. 2.3.8. It can be observed that the MOS degrades as the microphone-to-speaker distance increases, however the degradation in MOS is very small for these conditions. Similarly for the coloration conditions, the perceptual difference between the three shelf filters is less than 0.9 MOS. Figure. 2.3.9 shows the condition averaged MOS obtained for the MNRU conditions, where again a compression effect is observed for the various levels of the Q factor. Figure. 2.3.10 presents a histogram of the MOS for the C-Qual database, as can be seen, the region between 3 and 4 MOS has the greatest number of observations.

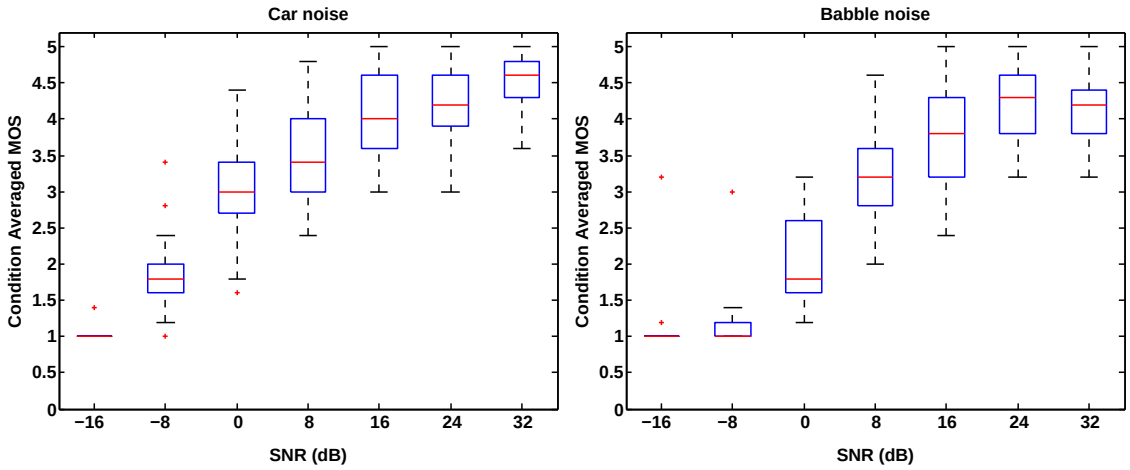

Figure 2.3.4: Relationship between SNR and subjective MOS for car (left) and babble (right) noise. Five outliers are detected over the -16 to 32 dB SNR range.

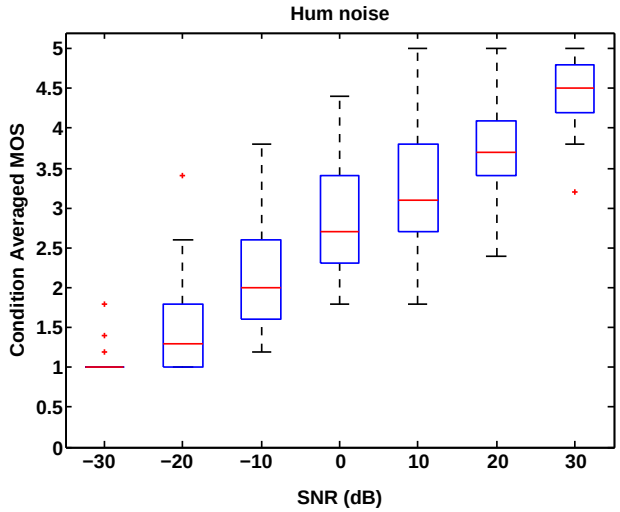

Figure 2.3.5: Relationship between subjective MOS and SNR for hum noise. Three outliers are detected for the -30 dB SNR condition.

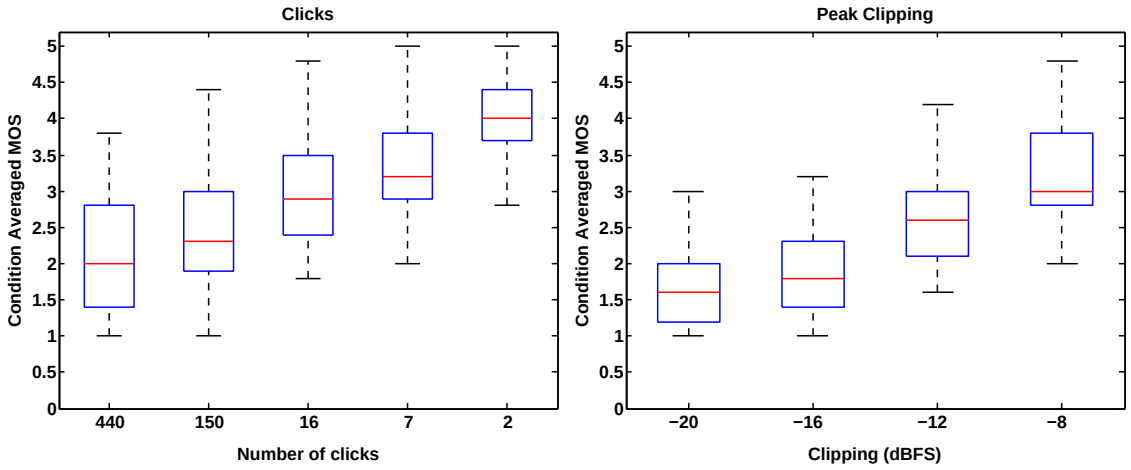

Figure 2.3.6: Relationship between the number of clicks (left) and peak clipping (right) and subjective MOS.

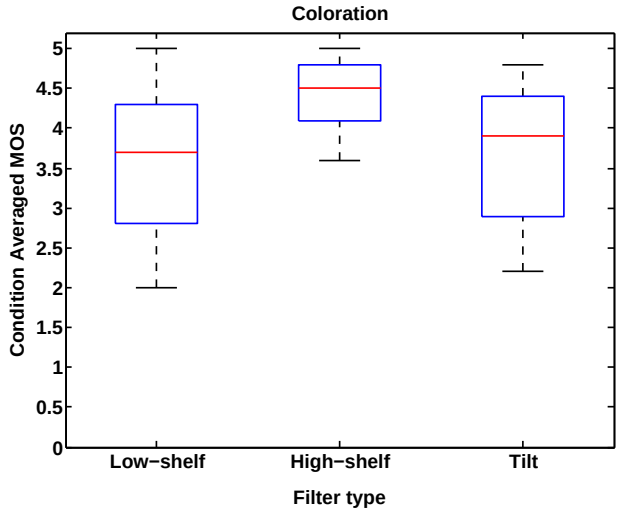

Figure 2.3.7: Subjective MOS for the three shelf filters tested in C-Qual.

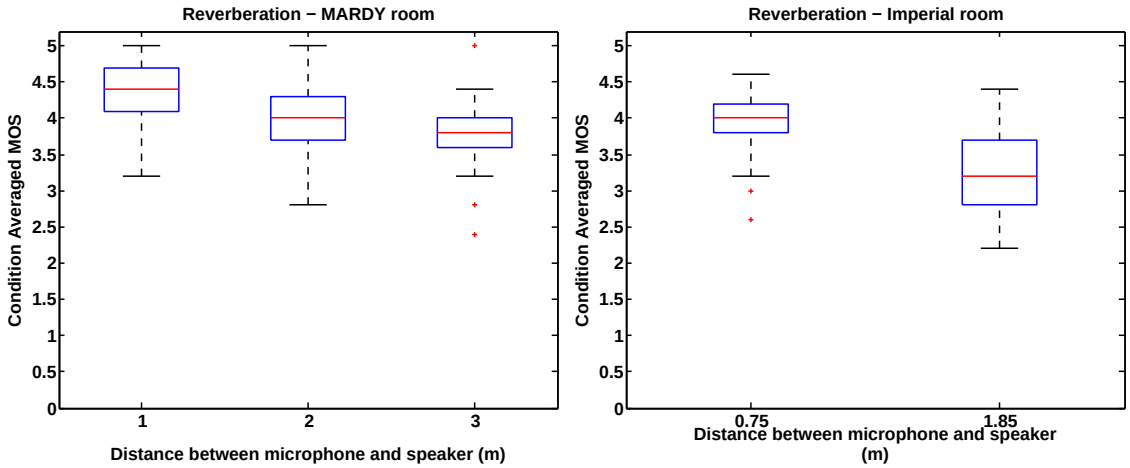

Figure 2.3.8: Subjective MOS for reverberation with three microphone to speaker distances for the MARDY room (left) and Imperial room (right).

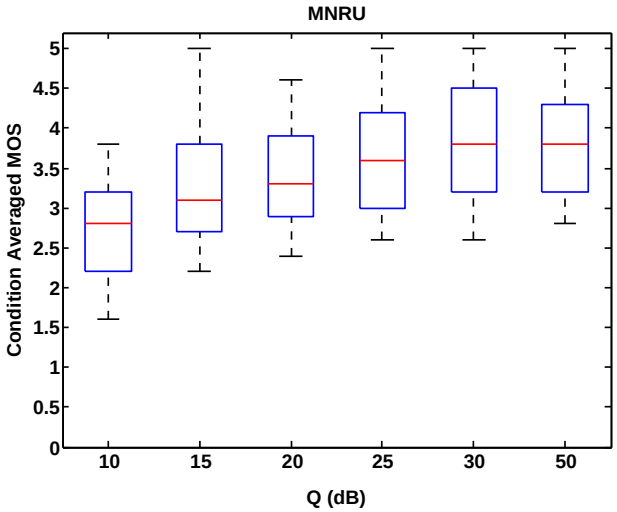

Figure 2.3.9: The relationship between MNRU and MOS for the C-Qual database.

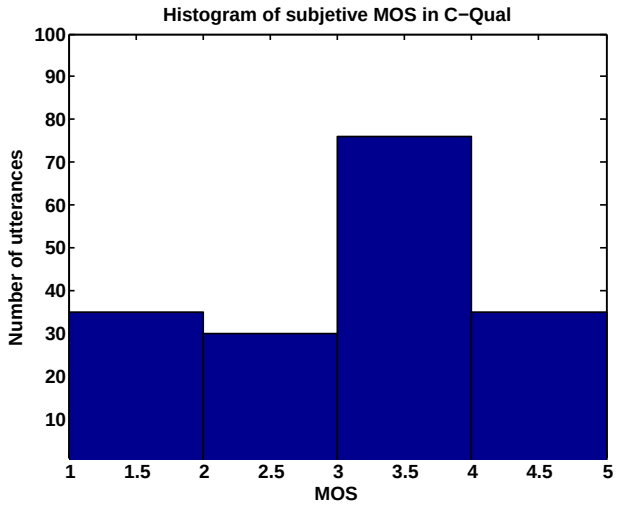

Figure 2.3.10: A histogram of the MOS obtained in the C-Qual database.

Reliability of data

PCA was used for assessing the inter-subject and the intra-subject reliabilities [168]. A high inter-subject reliability of 0.91 indicated that the subjects gave consistent responses in each of the five sessions. Furthermore, a high intra-subject reliability of 0.93 indicated that the naïve listeners gave reliable responses. In comparison, the P.23 database was found to have an intra-subject reliability of 0.89.

2.3.2 Extension to additional speech material

As the performance of a machine learning algorithm also depends on the quality and quantity of training data [92], the development of a large corpus of human labelled training data is an important issue. However, the use of human subjects for labeling data for a particular application such as speech quality is an expensive and time consuming task [139]. If the C-Qual database (described in Section 2.3.1) was split equally into a training and test partition, each partition would consist of only 84 audio examples. In order to develop the speech assessment algorithms a larger variety and quantity of data is required, for example the training database for the LCQA algorithm contains 6318 speech utterances [53]. In the law enforcement scenario, such extensive databases are not available and the use of subjective testing for developing extensive training databases is not feasible.

A possible solution is to use an intrusive objective technique to perform the labeling as this would allow a large amount of training data to be automatically labelled. It is necessary however to ensure that the degradations that are being labelled are relevant for the intrusive algorithm. In the case of speech quality assessment, the C-Qual data can be used to perform a validation of the intrusive method and utilize it only for those degradations that are well predicted. The same can also be done for speech intelligibility assessment by utilizing the subjective study in [66] for example. The automatic labeling method is outlined as follows

- Select a robust intrusive technique for the assessment criteria, such as speech quality.
- Validate the performance of the intrusive method on a subjective test database.
- Generate a training database with the degradations for which the intrusive method achieves a good performance.

The databases and training algorithm for each speech assessment criteria are described in the respective chapters. The next section will outline the NISA framework.

2.4 NISA Framework

The NISA framework is a data-driven machine learning approach to speech assessment that uses a combination of feature extraction followed by a tree based model. Machine learning based approaches have been applied successfully to many complex problems [31] including non-intrusive speech quality [180, 36, 53] and intelligibility [145] estimation as well as CODEC [149] detection and verification. Whereas this overall approach is commonly adopted in the literature, the novelty in the NISA framework lies both in the way NISA extracts features more discriminative in the given task, and in the way it performs the modelling of these features for which it employs a CART approach. The overall framework is shown in Fig. 2.4.1 and a description of each component of the framework is presented in the following subsections.

2.4.1 Pre-processing

The first step is the short-time segmentation of the input signal $y(n)$ into 20 ms frames by applying a non-overlapping Hanning window. The resulting signal is denoted as $y(i)$, where i is a 20 ms frame. The second step is application of a Voice Activity Detector (VAD) based on the P.56 method [82] to select frames where speech is present. The VAD is a basic energy based method that first computes the speech level of the entire signal using the P.56 [82] method and selects those frames that have a speech level within a range dependent on the P.56 level.

The next step is a normalisation of the energy in the speech active frames, this is done to make the feature extraction that follows to be gain independent. This then followed by short-term feature extraction (Section 2.4.2). The statistics of the short-term features are used to characterise the entire signal and combined with long-term features (Section 2.4.3) to create the final feature vector, φ, for the current signal. The features, φ, are used to infer a trained CART model, $\mathbb{T}$, that has been previously trained on a feature matrix, Φ, with corresponding ground truth scores from a training database.

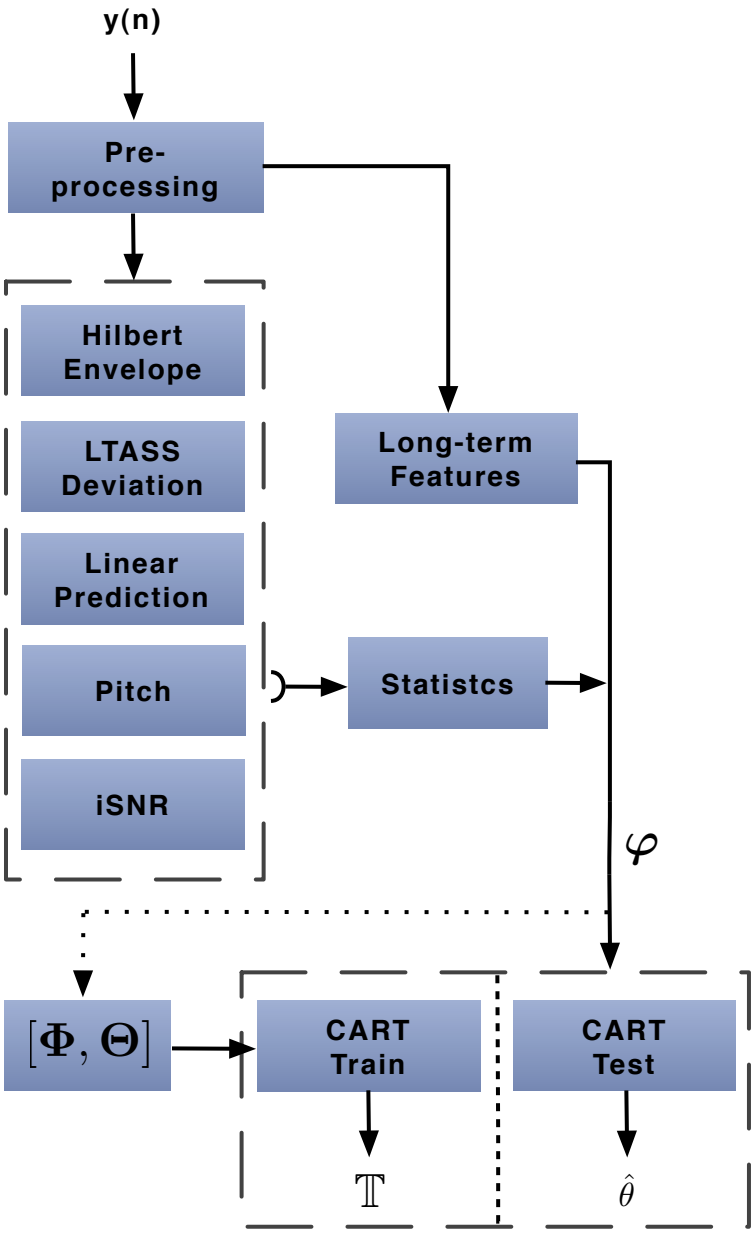

Figure 2.4.1: Block diagram of the NISA framework. The first stage is a pre-processing of the noisy speech signal, $y(n)$, followed by feature extraction. In the test mode, a pre-trained CART model is evaluated using the feature vector, φ, for the current signal, resulting in the estimated label, $\hat{\theta}$. In the training mode, the feature vector is combined into the training feature matrix, Φ, and used to construct a CART model.

2.4.2 Short-time features

The NISA framework computes the mean, variance, skewness and kurtosis of the short-time features over the entire speech utterance[6]. These are combined with other long-term features to form the final feature vector, φ, for the current utterance.

2.4.2.1 Pitch

A comparison of various pitch trackers was made in [147], since then the PEFAC algorithm [52] has been developed and has been shown to be robust to high levels of additive noise. The following subsection presents a brief validation of four pitch estimation techniques on the additive noise conditions of the C-Qual database.

Pitch estimation in additive noise

The aim is to validate the performance of four pitch tracking algorithms on the additive noises of the C-Qual database and speech from the SAM [25] database. The SAM database contains 2 male and 2 female speakers with contemporaneous recordings of EGG signals. The SIGMA [161] algorithm (Matlab implementation from [21]) was used for the extraction of GCI from the EGG signals and the pitch period was defined as the time between two GCIs. Then the pitch period was interpolated into 20 ms frames for evaluating the pitch estimation algorithms, which provide pitch estimates in short-time frames. The analysis was restricted to the regions of the signal where pitch was present (as obtained from SIGMA). The overall measure of performance is referred to as the Modified Hit Rate (MHR), which is the percentage of pitch frames that have a pitch estimate with an accuracy (absolute difference between the estimated and true pitch per frame) of 80% or higher (see [147] for further details). The following four pitch estimation algorithms were tested, PEFAC [52] (Matlab implementation from [21]), RAPT [159] (Matlab implementation from [21]), YIN [28] (C implementation from original author) and an autocorrelation based pitch algorithm used in [53] (lpcauto.m implementation from [62]), referred here as the AC method. The overall results for pitch estimation, aggregated over all the noise types and SNR is shown in Table. 2.1, where PEFAC can be seen to be significantly better than the other methods. The PEFAC algorithm performs particularly well at low SNR

[6]Here an 'utterance' is defined as a segment of speech for which the measure of interest is constant. The duration of an utterance should be suitably long as to permit estimation of the various features to be employed. In this thesis the typical utterance duration is in the range 3 to 8 seconds. Long speech segments with varying quality or intelligibility can, without loss of generality, be segmented into shorter segments with less variability in the measure of interest.

conditions as illustrated by Fig. 2.4.2 and also in hum noise (Fig. 2.4.3). Appendix B presents histograms of the error in pitch estimation for the four algorithms presented here.

Algorithm	MHR(%)
PEFAC	73.4
YIN	46.1
RAPT	44.5
AC	16.8

Table 2.1: Overall performance of the four pitch estimation algorithms on the additive noise partition of the C-Qual database.

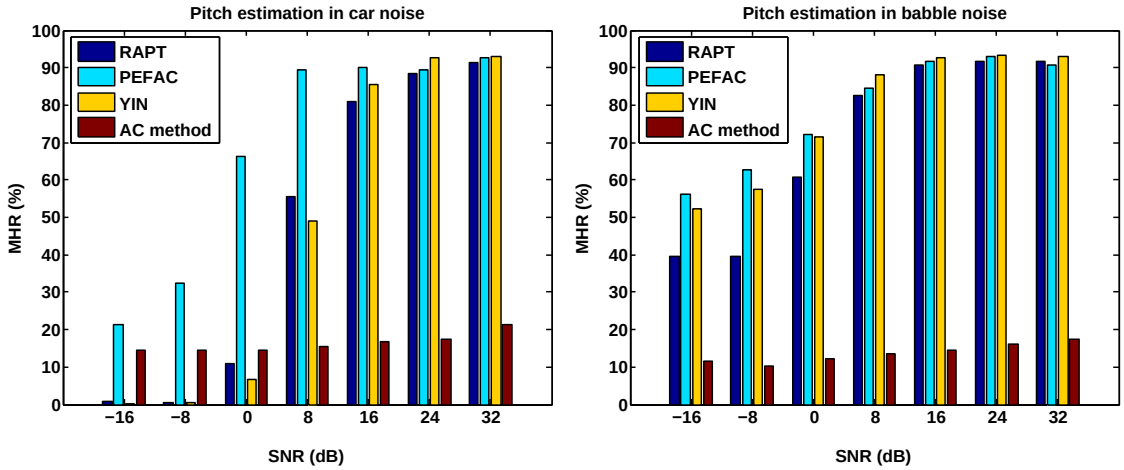

Figure 2.4.2: Pitch performance for car (left) and babble noise (right).

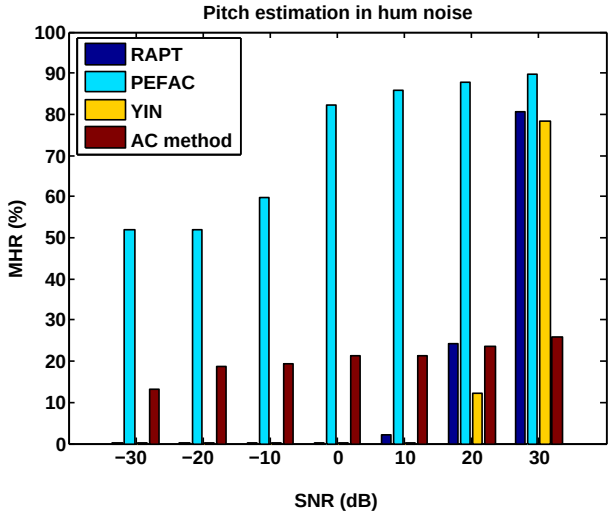

Figure 2.4.3: Pitch performance for hum noise.

2.4.2.2 Importance weighted Signal to Noise Ratio (iSNR)

The SNR of a speech signal is an objective measure of the relative level of distortion in the output signal and can be defined as the ratio of the speech power to the noise power as follows

$$\text{SNR} = 10 \times \log_{10}(\frac{P_s}{P_V}), \tag{2.4.1}$$

where P_s is the speech power and P_V is the noise power and the following additive model for the noise signal is assumed

$$y(n) = s(n) + v(n)$$

where $y(n)$ is the noisy speech signal, $s(n)$ the clean speech signal and $v(n)$ is the noise signal. The SNR definition presented in 2.4.1 is an intrusive measure where the noise and speech power is known. The iSNR feature presented here is a non-intrusive SNR measure that performs the SNR calculation in short-time frames and also applies a frequency weighting function based on speech intelligibility measurement. The iSNR feature uses the 1/3 octave frequency band importance function from the SII standard ([145, 8]). This function applies more weight to the signal at the frequencies that have a higher importance to speech intelligibility as shown in

Figure 2.4.4. The iSNR for time frame i is defined as:

$$\text{iSNR}(i) = 10 \times \sum_{k=1}^{N_k} I(k) \times log_{10}\left(\frac{max(0, P_y(i,k) - P_{\tilde{v}}(i,k))}{P_{\tilde{v}}(i,k)}\right) \tag{2.4.2}$$

where $I(k)$ is the SII weighting function, N_k is the number of frequency bands, $P_{\tilde{v}}(i,k)$ is the estimated noise power spectrum obtained by the minimum statistics algorithm [113, 21] and $P_y(i,k)$ is the power spectrum of the noisy speech signal, calculated as:

$$P_y(i,k) = Y(i,k) \times Y^{\star}(i,k) \tag{2.4.3}$$

where $Y(i,k)$ is the Discrete Fourier Transform (DFT) of the noisy signal[7]. Additionally, the rate of change of the iSNR feature over all voiced frames is computed.

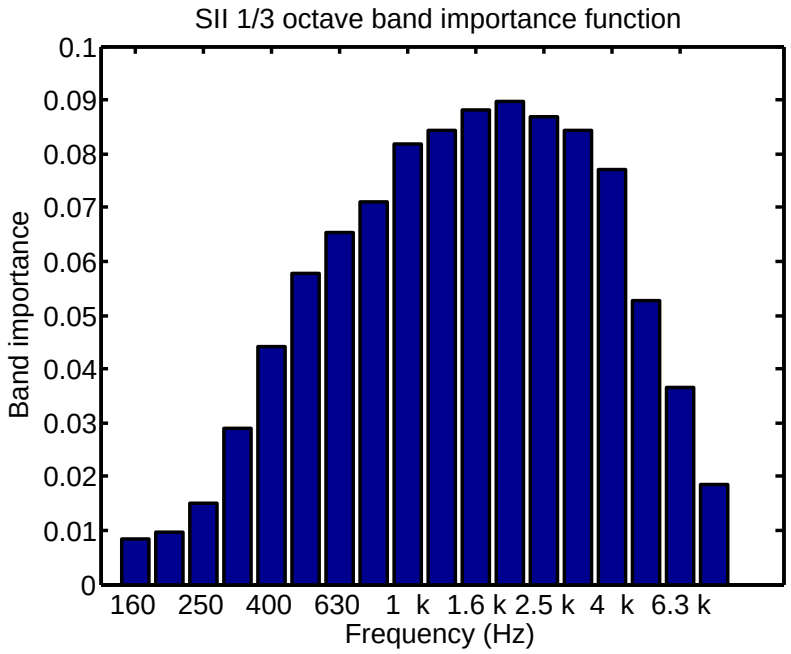

Figure 2.4.4: The 1/3rd octave frequency band importance function from the SII standard (Table 3).

2.4.2.3 Hilbert envelope

The Hilbert decomposition of a signal results in a slowly varying envelope and a rapidly varying fine time structure component. The envelope of a speech signal, obtained through Hilbert

[7] In practice, the FFT is used.

decomposition is an important factor in speech reception [153]. The envelope for frame i of a signal is calculated as:

$$e_i = \sqrt{y(i)^2 + |\mathscr{H}(y(i))|}, \tag{2.4.4}$$

where e_i is the envelope of the i^{th} frame of $y(n)$ and $\mathscr{H}\{.\}$ is the Hilbert transform. The Hilbert transform of a time domain signal, $x(n)$, can be expressed as the response of a linear filter with impulse response, $(\pi n)^{-1}$, as follows [14]

$$\mathscr{H}(x(n)) = \frac{1}{\pi n} \otimes x(n) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x(\tau)}{n-\tau} d\tau,$$

where $\otimes$ is the convolution operator. The following two features are extracted from the per-frame signal Hilbert envelope. The variance (σ_{e_i}) and dynamic range (Δ_{e_i}) of the per-frame envelope are computed as follows:

$$\sigma_{e_i} = \frac{1}{N_i} \sum_{i=1}^{N_i} (e_i - \mu_{e_i})^2 \tag{2.4.5}$$

$$\Delta_{e_i} = |\max(e_i) - \min(e_i)|. \tag{2.4.6}$$

Additionally, the rates of change of these features over all frames are also included.

2.4.2.4 LTASS deviation

The Long Term Average Speech Spectrum (LTASS) has a characteristic shape that is often used as a model for the clean speech spectrum and has been used in a number of speech processing algorithms, such as blind channel identification [49]. The ITU-T P.50 [84] standard defines an analytic expression for approximating LTASS as shown in Fig. 2.4.5. The PLD for frame i and frequency bin k is defined as:

$$\text{PLD}(i,k) = \log(P_y(i,k)) - \log(P_{LTASS}(k)), \tag{2.4.7}$$

where $P_y(i,k)$ is the magnitude power spectrum of noisy signal and $P_{LTASS}(k)$ is the LTASS power spectrum. This deviation spectrum measures the effects on the magnitude spectrum due to the distortion. The per-frame LTASS deviation spectrum is used to derive the spectral flatness (SF), spectral centroid (SC) and spectral dynamics (SD) features as defined below:

$$SF(i) = \frac{\exp\left(\frac{1}{N_k}\sum_{k=1}^{N_k}\log(\mathrm{PLD}(i,k))\right)}{\frac{1}{N_k}\sum_{k=1}^{N_k}\mathrm{PLD}(i,k)}, \tag{2.4.8}$$

$$SC(i) = \frac{\sum_{k=1}^{N_k}\omega(k)\times\log(\mathrm{PLD}(i,k))}{\sum_{k=1}^{N_k}\log(\mathrm{PLD}(i,k))}, \tag{2.4.9}$$

$$SD(i) = \frac{1}{N_k}\sum_{k=1}^{N_k}(\log(\mathrm{PLD}(i,k) - \log(\mathrm{PLD}(i,k)))^2, \tag{2.4.10}$$

where ω is a frequency index vector and N_k is the number of FFT bins. The spectral flatness, dynamics and centroid of LTASS deviation spectrum and their rate of change are included as short-term features.

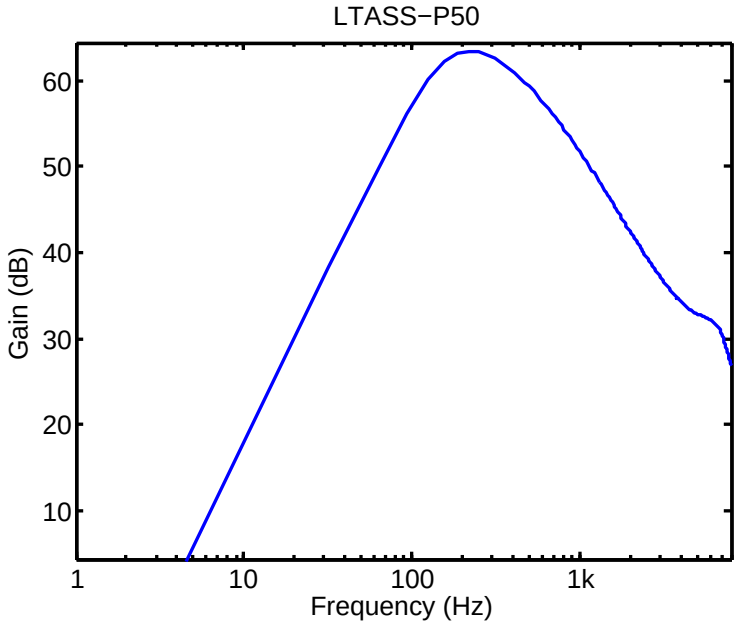

Figure 2.4.5: Long-term average speech spectrum (LTASS) from the ITU-T P.50 standard.

2.4.2.5 LPC features

A 10th order linear predictive coding (LPC) is performed on the speech signal using the auto-correlation method [132]. The residual variance and its rate of change over the utterance are included as features. Additionally, the spectral centroid, flatness and dynamics of the magnitude response of the LPC spectrum and their rate of change are computed, as in [53], and used in addition.

2.4.3 Long-term features

The long-term features are based on the deviation of the long-term spectrum of the signal from LTASS, as described in the following subsection. This differs from the PLD based features, which are computed per time frame of the signal.

2.4.3.1 LTASS deviation

The long-term deviation of the magnitude spectrum of the signal (calculated over the entire utterance) is defined as follows

$$P_{LTLD}(k) = \frac{1}{N_i} \sum_{i=1}^{N_i} \mathrm{PLD}(i,k) \tag{2.4.11}$$

where k if the frequency index, PLD is the power spectrum of long-term deviation (2.4.7). It is expected that this feature could help identify the long-term frequency characteristics of different types of degradations. Figure. 2.4.6 shows a plot of the P_{LTLD} feature for speech from a female speaker degraded with car and babble noise at -16 dB SNR, where it can be observed that the effects of the two noise types on the long-term spectrum of the speech signal can be identified.

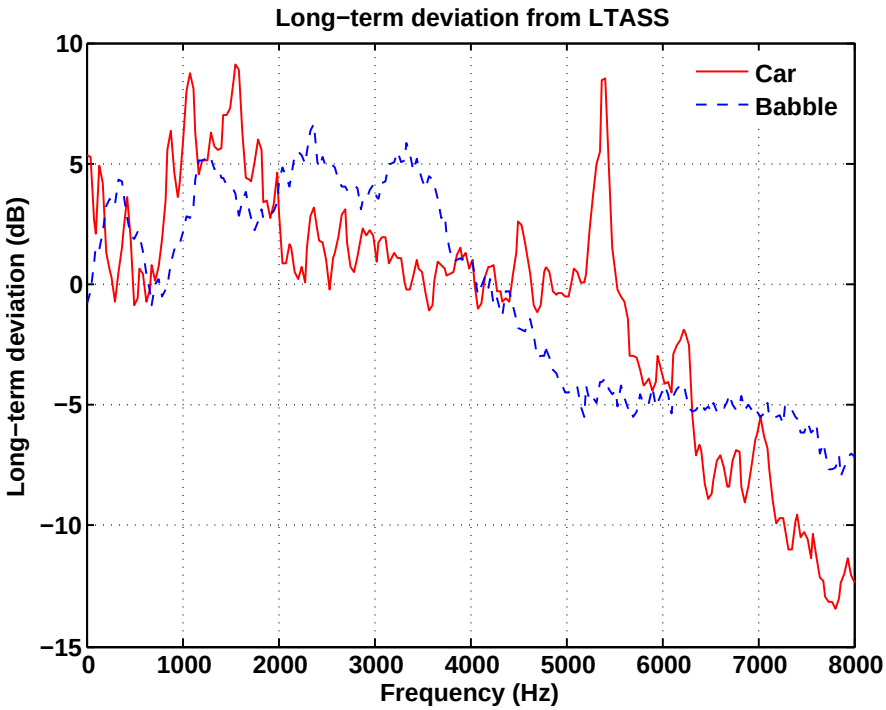

Figure 2.4.6: The long-term deviation feature, P_{LTLD}, for car and babble noise at -16 dB SNR. The speech is a pair of utterances from a female speaker from the C-Qual database.

2.5 Summary

This chapter presented a review of the machine learning and speech analysis technology that provided the foundation on which to develop the non-intrusive speech assessment framework. This included a review of the CART machine learning algorithm, feature dimensionality reduction methods, linear prediction and pitch estimation. The C-Qual database was presented in Section 2.3.1, which contains degradations reflective of the law enforcement scenario, labelled with subjective speech quality scores. The results for the 44 degradation conditions was presented and confirmed to be of high reliability.

Since the performance of a data-driven technique is affected by the quality and quantity of training data and large scale subjective testing is an expensive task, a method for automatically generating training data was proposed in Section 2.3.2. Finally, the data-driven NISA framework for speech assessment was presented in Section 2.4. The framework is based on a CART model and a two stage feature extraction that models the statistics of short-term features and also uses a number of long-term features. A number of novel features for speech assessment were also presented. The NISA framework will be applied in the Chapter 3 for speech quality assessment and in Chapter 4 for speech intelligibility assessment. Also, the NISA framework will be applied to non-intrusive CODEC identification and verification in Chapter 5.

Chapter 3

Speech Quality Assessment

THIS chapter presents the problem of speech quality assessment, beginning with an introduction of the concept to speech quality and then reviewing current methods for assessing speech quality. The problem of per-utterance speech quality estimation is presented in Section 3.4, followed by an initial study on time-varying objective quality assessment in Section 3.5. The research presented here relates in part to the following publications [146, 148].

3.1 Introduction

Speech quality is a judgement of a perceived multidimensional construct that is internal to the listener and is typically considered as a mapping between the desired and observed features [119] of the speech signal. Speech quality assessment can be used for analyzing the perceptual effects of various degradations on a speech signal. These degradations are caused when telecommunications and surveillance systems are deployed in non-ideal operating conditions and the problem is compounded further by the increasing complexity and non-linear processing integrated into modern communication systems [136]. In the telecommunications industry, such degradations impact the Quality of Service (QoS) of a system and objective techniques for speech quality assessment can be used for optimizing network parameters, capacity management and cost optimization based on customer experience [136, 139, 53]. In the law enforcement context, degradations can effect transcription rate and accuracy [125] and speech quality assessment can be used for segmentation of long recordings, allowing an audio analyst to identify sections of usable quality speech.

3.2 Review

The quality of a speech signal can be assessed in a subjective experiment by asking a number of human subjects to provide judgements of the signal quality or estimated using objective techniques. The estimation task may be restricted to short speech utterances or the time-varying assessment of the signal quality. This section reviews the current literature on speech quality estimation using the per-utterance model (Section 3.2.1) and the time-varying model (Section 3.2.2).

3.2.1 Per-utterance assessment

Current methods for speech quality assessment are focused on a short speech utterance model, where the quality of short speech utterances (typically between 3 and 8 seconds in length) are degraded by uniform degradations (in many single degradations are used) and a single rating is estimated for a test signal [61, 75]. Moreover, a large focus is on listening-only methods, where the effects in a listening scenario are considered, whereas conversational quality tests aim to evaluate the quality in a conversational environment where there is an interaction between two subjects [136, 139]. The following subsection describes various techniques for short-time (utterance level) listening-only speech quality assessment, beginning with subjective measurement, followed by objective techniques.

Subjective assessment

As the quality of a speech signal is a highly subjective measure [53], a number of techniques for subjective speech quality assessment have been proposed. The International Telecommunication Union (ITU) standard [75] outlines a number of protocols for carrying out subjective quality experiments on various measurement scales. There are broadly two types of subjective tests, one where the subjects rate the absolute quality of a signal (absolute rating) and the other where subjects provide a preference for one of a pair of signals (preference rating). A frequently used rating scale for absolute rating is the 5-point ACR listening quality scale (Table 3.1).

The ACR scale can have a low sensitivity in discriminating between signals with high quality and in such situations, the Degredation Category Rating (DCR) is often used (Table 3.1) [75]. Table 3.2 shows the recommended 7-point Comparison Category Rating (CCR) scale [75] for obtaining preference ratings. The quality scores obtained from a subjective experiment is re-

	ACR	DCR
Rating	Quality of the speech	Description of the degradation
5	Excellent	Inaudible
4	Good	Audible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 3.1: The 5-point ACR and DCR absolute rating scales.

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

Table 3.2: The 7-point CCR preference rating scale.

ferred as Mean Opinion Score for Subjective Listening Quality (MOS-LQS). Although it is possible to get accurate results with subjective testing for small quantities of data (and are believed to give the "true" speech quality), they are time consuming and expensive to administer for large amounts of audio and thus unsuitable for real-time (or even near real-time) applications [42]. The objective methods for speech quality assessment aim to overcome these issues by modeling the relationship between the desired and perceived characteristics of the signal algorithmically, without the use of listeners. These objective methods for speech quality assessment are presented in the following subsection.

Objective assessment

There are three main categories of objective speech quality assessment, those which require a reference (unprocessed) signal in addition to the received (processed) signal are referred to as intrusive techniques, those that rely only on the received signal are referred to as non-intrusive techniques and those that rely on the parameters of the processing system are commonly referred to as parametric techniques [41] (Fig. 3.2.1). The quality score estimated with an intrusive or non-intrusive technique is referred as Mean Opinion Score for Objective Listening Quality (MOS-LQO) and when a parametric method is used, it is known as Mean Opinion Score Estimated with a Parametric Listening Quality algorithm (MOS-LQE). The parametric methods

estimate speech quality by measuring various properties of the transmission system under test and require a full characterization of the system [139]. The E-model [81] is considered an archetypal parametric model [119] that is useful as a transmission planning tool for optimizing the transmission system parameters. The result of the E-model is a transmission rating factor that can be transformed to a MOS scale [81]. The parametric methods can be used in situations where the system parameters are available, but when these are not known a signal based approach must be adopted.

The simplest signal based objective methods include the SNR and segmental-SNR, which are of a low computational complexity [53]. The SNR metric is widely used to assess speech quality for speaker identification [15], however the SNR metric can have a poor correlation with subjective quality scores if different types of distortions are compared [53].

Intrusive methods are used where access to a clean signal is possible, such as CODEC development or for assessing the quality of a communication system with known test signals. An ITU industry standard for intrusive quality testing is the PESQ [85] measure, which is an integration of two previous intrusive methods [137]: an extended version of Perceptual Speech Quality Measure (PSQM) [13] and the Perceptual Analysis Measurement System (PAMS) [140, 138]. The PESQ algorithm has been extended for the assessment of wide-band telephone networks and speech CODECs and standardized as Wide-band PESQ [88]. In PESQ, quality scores are determined on a scale from -0.5 to 4.5 and a mapping function is then used to map the PESQ score to mean opinion scores (MOS) [86]. A correlation coefficient of 0.935 between PESQ MOS-LQO (mapped with the function in [137]) and MOS-LQS has been reported for a number of telecommunication relevant databases [139, 137]. More recently, an extension of PESQ has been standardized as POLQA [89].

In situations where the clean signal is not available, a non-intrusive technique may be applied. A number of non-intrusive techniques have been proposed over the past decade, see [139, 119] for a review. The current ITU-T industry standard algorithm for non-intrusive speech quality assessment is the P.563 [80], which uses a number of features from the audio stream to estimate the quality score directly on the MOS scale. This method was chosen from an ITU-T competition between 2002 and 2004 and is the result of a collaboration between three companies specializing in speech quality assessment [111], known as the Single-Ended Assessment Model (SEAM), which beat a competing method [139] known as ANIQUE [98]. The ANIQUE+ has since been standardized as an American National Standard Institute (ANSI) standard [6].

More recently, a number of data-driven methods have been proposed that derive a number of features from the speech signal and use a previously trained model to map the features to a

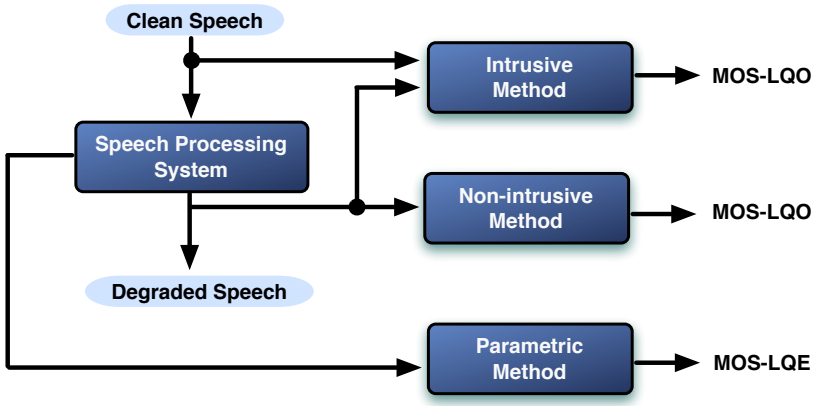

Figure 3.2.1: Intrusive and non-intrusive objective speech assessment techniques.

quality score. A number of techniques that use machine learning models such as GMMs to model perceptual speech features such as the Perceptual Linear Prediction (PLP) [63] coefficients have been proposed by Falk *et al.* [40, 39, 36, 38]. Additionally, speech quality measures based on a data-mining approach using CART have also been developed [37, 179, 180]. The LCQA algorithm [53] derives a number of features from the speech signal and has been shown to outperform the P.563 measure for a large set of degradations and due to its low complexity, novel feature set and favorable performance this method is selected for baseline comparison along with the P.563 method. Further details of these are presented in Section 3.3.

3.2.2 Time-varying assessment

Over the last two decades, a number of methods for subjectively measuring the quality of speech have been developed [75], typically using short speech sentences (between 3 and 8 seconds in length) each degraded by stationary degradations. A number of databases are available for speech quality research using the P.800 protocol [146]. Objective methods have also been developed and validated for providing estimates of speech quality for short sentences of speech with homogenous degradations. However, in realistic scenarios, both for mobile telecommunication devices and in law enforcement applications degradations have an inherently time-varying nature. Also, the duration of an average communication is much longer than the standard 8 seconds considered in typical subjective quality tests.

Previous studies of subjective measurement of time-varying speech quality include Hansen *et al.* [61] which used the modulated noise reference unit (MNRU) to measure the quality of isolated words (durations from 0.135 s to 0.911 s) as well as continuous quality assessment

using two different SNR profiles on 40 s of ongoing speech. It is reported that subjects can assess the quality of words in isolation as an instantaneous task and reliably assess the time-varying quality of continuous speech with a delay of 0.5 s. Similar studies have been reported by Voran *et al.* [171] and Heute *et al.* [65]. A protocol for subjective measurement of continuous, time-varying speech quality [79] has now been standardized. The protocol describes the interface for collecting time-varying quality scores using a potentiometer and a graphical user interface for collecting the overall quality of the stimuli. A duration of 40 seconds to 3 minutes for the stimuli is recommended for time-varying speech quality assessment.

3.3 Current methods

3.3.1 PESQ

PESQ is one of the most widespread intrusive speech quality assessment algorithm [119] and is used in this thesis as a ground truth estimator of speech quality for automatically labeling large quantities of training and test data. This section provides an overview of the PESQ algorithm and a validation of the method is then presented for the C-Qual database in Section 3.4.4.

The main elements of the PESQ algorithm are presented in Fig. 3.3.1 [85, 137], beginning with the clean and degraded speech signals at the top and resulting in the estimated PESQ score at the bottom. The first stage is a pre-processing of the clean and degraded signals, where first a level alignment is performed to set the signal power to a constant level (calculated assuming that the subjective listening level is 79 dB SPL [85, 78]), followed by an Intermediate Reference System (IRS) receive filter to model a standard telephone handset. This is then followed by the the following processing modules:

- Time-alignment - this module computes any piecewise constant delays between the clean and degraded signal for accurate comparison of the two signals in the auditory transform and time integration modules. The output of this stage is the delay per time interval (d_i).
- Auditory transform - a psychoacoustic model based on the Bark spectrum is applied, where the signals are segmented into 32 ms frames with a Hamming window (50% overlap between frames) and transformed to the frequency domain by an FFT and mapped to the pitch scale using a modified Bark scale [137]. Then a linear frequency equalization is performed on the clean signal by calculating the frequency transfer function between the degraded and clean signal. This is followed by an equalization of gain variations between the two signals. Then the intensity representation of the signals is mapped to a scale of the perceived loudness in time and frequency.
- Perceptual difference - the absolute difference between the loudness densities of the clean and degraded signal is calculated and represents the audible error, referred here as the residual spectrum. A threshold in each time-frequency bin of the residual spectrum models the effects of masking of small distortions in the presence of loud signals. The resulting signal is the symmetric disturbance density.
- Asymmetry processing - the output of this module is the weighted asymmetric disturbance that measures additive distortions. The asymmetry factor is the ratio of the degraded

and clean signal densities in each time-frequency bin, raised to the power of 1.2. The factor is bounded with an upper limit of 12.0 and values of the asymmetry factor less than 3 are set to 0. The resulting signal is the asymmetric disturbance density.

- Frequency and time integration - the symmetric and asymmetric disturbance densities are integrated over all frequency bins for each time frame using L_p norms and weighting function that emphases distortions occurring in silent segments of the speech. The result is the frame densities which are aggregated over intervals of 20 frames using L_6 norms and multiplied by a recency factor and summed over all time frames [85]. The result is the symmetric disturbance and asymmetric disturbance factors (d_{SYM} and d_{ASYM} respectively).

- Disturbance mapping - the final PESQ score is then calculated as

$$\hat{\theta} = 4.5 + (\alpha \times d_{SYM}) + (\beta \times d_{ASYM}),$$

where $\hat{\theta}$ is the PESQ score, d_{SYM} is the symmetric disturbance factor and d_{ASYM} is the asymmetric disturbance factor. The values of α and β were determined by analysis using a database of 30 subjective tests as -0.1 and -0.0309 respectively [137]. The result is the PESQ score with a value in the range [-0.5 to 4.5], although for normal subjective test material a lower limit of 1.0 is observed [137]. Furthermore, a mapping function from PESQ score to MOS-LQO is commonly used to allow a comparison with subjective quality scores on the same scale and an ITU standard outlines a reversible mapping function for this purpose [86].

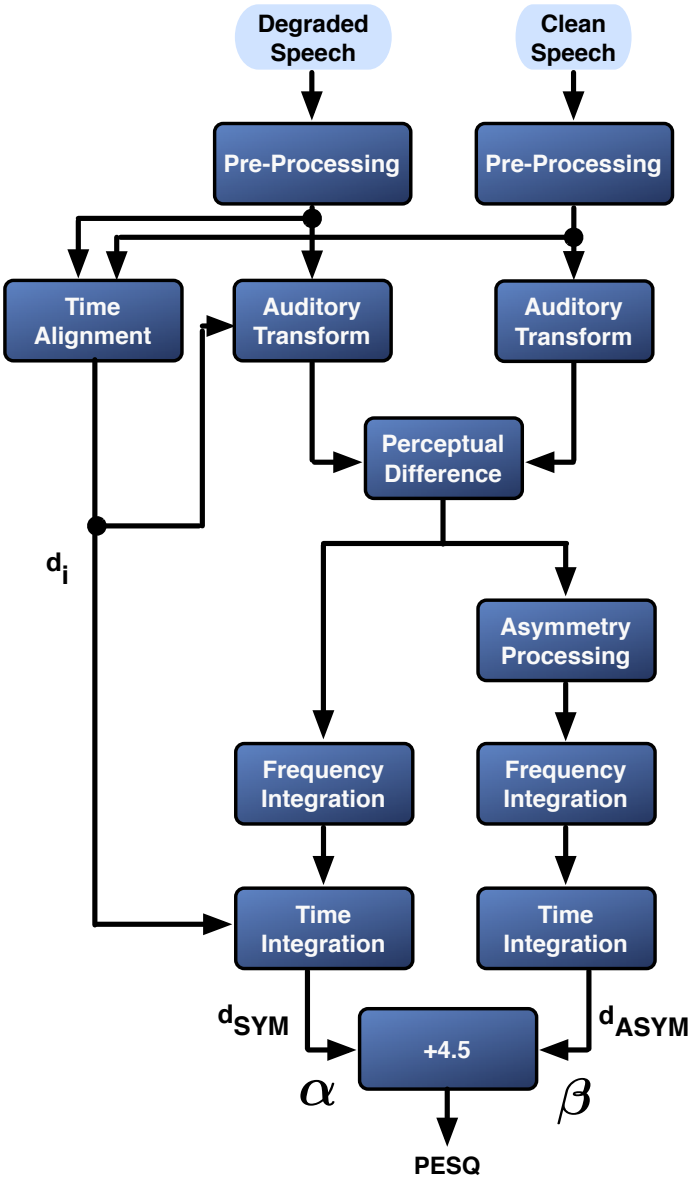

Figure 3.3.1: PESQ algorithm overview

3.3.2 P.563

The ITU standard algorithm for non-intrusive speech quality assessment is the P.563 [80]. There are three main aspects to the calculation of the quality score, the first one is the analysis of variations in the voice production system, the second aspect is a reconstruction of a clean reference signal and apply an intrusive perceptual model to evaluate speech quality and finally, a number of distortion specific parameters are calculated and a quality score is estimated by a linear combination of these parameters [111] as shown in Fig. 3.3.2. The key modules in the P.563 method are summarized as follows [111]:

- Pre-processing : the first step is a level normalization of the input signal to -26 dB Overload (dBOV) followed by input filtering and voice activity detection.
- Vocal tract model and LPC analysis : this module aims at analyzing the voice production system for discrepancies from an expected 'ideal' model to identify distorted speech. The first stage is pitch estimation and refinement using a hybrid temporal/spectral method, followed by voiced/unvoiced classification. The method of deriving vocal tract models from voiced speech for quality assessment from [54] is used to provide an estimate of speech distortion. Additionally, cepstral and linear prediction (21 order) coefficients are extracted for each voiced speech frame and their statistics (skewness and kurtosis) are calculated and compared with values obtained from clean speech [80]. This provides further information on the unnaturalness of the signal.
- Speech reconstruction and intrusive modeling : this model begins with an estimation of a quasi-clean speech signal using LPC analysis, vocal tract constraint and LPC synthesis. The estimated clean signal is then used in a perceptual model similar to the one used in PESQ to compute an additional estimate of the speech quality [111].
- Distortion parameters : the P.563 method also calculates a number of degradation specific parameters. These include additive noise characterization using an estimate of the SNR, robotization detection by computing cross-correlations of adjacent short-time frames, temporal clipping detection by analysis of the variations in the signal envelope and signal correlated noise detection.
- Distortion classification : the final step in calculating the estimated quality score is a linear perceptual weighting, guided by the dominant distortion identified previously. The annoyance order is shown in Table 3.3.

Annoyance order	Distortion class
1	High level background noise
2	Signal interruptions
3	Signal correlated noise
4	Speech robotization
5	Common unnaturalness

Table 3.3: Annoyance order for P.563 distortion classes [111].

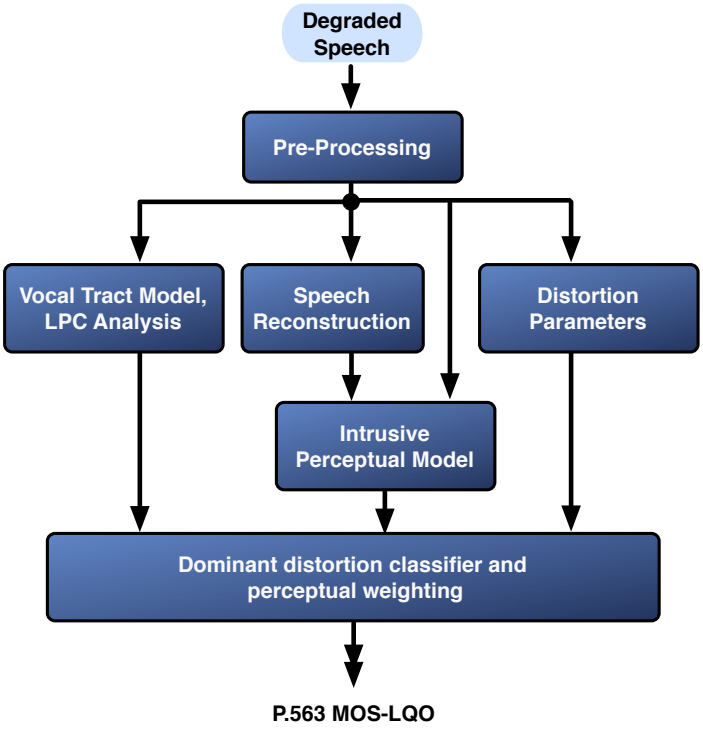

Figure 3.3.2: P.563 algorithm structure.

3.3.3 LCQA

The LCQA method is a machine learning approach to non-intrusive speech quality assessment and has been shown to outperform the P.563 method for a number of speech databases, achieving an average (per-condition) correlation of 0.94 on 7 databases (compared with 0.87 for P.563) [53]. The algorithm is outlined in Fig. 3.3.3 and begins with a pre-processing stage that splits the input signal into 20 ms non-overlapping frames for further processing. The remaining aspects of the algorithm are summarized as follows.

Feature extraction

The algorithm extracts 11 features per frame (denoted as ϕ in Table 3.4). The pitch period is extracted by an autocorrelation based method [99] and the spectral features are derived from a 10^{th} order LPC analysis of the speech signal. The spectral flatness feature for time frame i is calculated as

$$\phi_1(i) = \frac{\exp\left(\frac{1}{N_k}\sum_{k=1}^{N_k}\log(P_{LPC}(i,k))\right)}{\frac{1}{N_k}\sum_{k=1}^{N_k}P_{LPC}(i,k)},$$

where $P_{LPC}(i,k)$ is the frequency response (frequency index k) of the LPC model magnitude spectrum, defined as

$$P_{LPC}(i,k) = \frac{1}{|1+\sum_{m=1}^{p}a_m e^{-jkm}|^2}.$$

Similarly, the spectral dynamics ($\phi_2(i)$) and spectral centroid ($\phi_3(i)$) features for the i^{th} time frame are calculated as

$$\phi_2(i) = \frac{1}{N_k}\sum_{k=1}^{N_k}(\log P_{LPC}(i,k) - \log(P_{LPC}(i,k)))^2,$$

$$\phi_3(i) = \frac{\sum_{k=1}^{N_k}\omega(k)\times\log(P_{LD}(i,k))}{\sum_{k=1}^{N_k}\log(P_{LD}(i,k))},$$

where $\omega(k)$ is the frequency vector[1].

[1] A vector containing the centre frequency of each FFT bin.

In addition to the 6 basic features, the rate of change of these over all time frames is also computed (Table 3.4). The next step is a frame selection procedure which applies thresholds on three per-frame features (ϕ_1, ϕ_2, ϕ_5) and retains only those frames that qualify this threshold. This is done to remove unnecessary[2] frames from the signal. This has been described as a generalization of a VAD and typically discards between 50% to 80% of the frames [53]. The new set of frames is denoted by $\tilde{\Omega}$.

Statistical description

The 11 per-frame features are described by their mean, variance, skewness and kurtosis as follows

$$\mu(\phi_j) = \frac{1}{N_{\tilde{\Omega}}} \sum_{i \in \tilde{\Omega}} \phi_j(i)),$$

$$\sigma(\phi_j) = \frac{1}{N_{\tilde{\Omega}}} \sum_{i \in \tilde{\Omega}} (\phi_j(i) - \mu(\phi_j))^2,$$

$$\gamma(\phi_j) = \frac{1}{N_{\tilde{\Omega}}} \frac{\sum_{i \in \tilde{\Omega}} (\phi_j(i) - \mu(\phi_j))^3}{\sigma^{3/2}(\phi_j)},$$

$$\kappa(\phi_j) = \frac{1}{N_{\tilde{\Omega}}} \frac{\sum_{i \in \tilde{\Omega}} (\phi_j(i) - \mu(\phi_j))^4}{\sigma^{2}(\phi_j)},$$

where ϕ_j is the j^{th} feature and $N_{\tilde{\Omega}}$ are the number of frames that are selected. The resulting 44 dimensional global feature vector (φ) is used to perform feature subset selection using the SFBS procedure on labelled training data. The resulting feature set[3] ($\hat{\varphi}$)is used for the GMM mapping stage.

[2] those frames that do not help improve the RMSE performance of the algorithm on the training data by a predetermined threshold.

[3] For the training data used in the LCQA paper, the final feature set contained 14 global features.

GMM mapping

The final quality estimate is obtained with a GMM mapping using final global features for the current signal and a trained GMM.

$$E(\theta|\hat{\varphi}) = \sum_{m=1}^{M} u^{(m)}(\hat{\varphi})\mu^{(m)}(\theta|\hat{\varphi}),$$

where

$$u^{(m)}(\hat{\varphi}) = \frac{w_m \times \mathcal{N}(\hat{\varphi}|\mu_{\hat{\varphi}}^{(m)}, \Sigma_{\hat{\varphi}\hat{\varphi}}^{(m)})}{\sum_{k=1}^{M} w_k \times \mathcal{N}(\hat{\varphi}|\mu_{\hat{\varphi}}^{(k)}, \Sigma_{\hat{\varphi}\hat{\varphi}}^{(k)})},$$

and

$$\mu^{(m)}(\theta|\hat{\varphi}) = \mu^{(m)}(\theta) + \Sigma_{\hat{\varphi}\theta}^{(m)}(\Sigma_{\hat{\varphi}\hat{\varphi}}^{(m)})^{-1}(\hat{\varphi} - \mu^{(m)}(\hat{\varphi})),$$

where $\mathcal{N}(\hat{\varphi}|\mu_{\hat{\varphi}}^{(m)}, \Sigma_{\hat{\varphi}\hat{\varphi}}^{(m)})$ is a multivariate Gaussian density and w is the mixture coefficient vector, $\mu^{(m)}(\theta)$ and $\mu^{(m)}(\hat{\varphi})$ are the means of the quality and feature vectors, $\Sigma_{\hat{\varphi}\hat{\varphi}}^{(m)}$ is the feature covariance matrix and $\Sigma_{\hat{\varphi}\theta}^{(m)}$ is the cross-covariance matrix of the m^{th} mixture.

Feature description	Feature	Rate of change of feature
Spectral flatness	ϕ_1	ϕ_7
Spectral dynamics	ϕ_2	-
Spectral centroid	ϕ_3	ϕ_8
Excitation variance	ϕ_4	ϕ_9
Speech variance	ϕ_5	ϕ_{10}
Pitch period	ϕ_6	ϕ_{11}

Table 3.4: The 11 per-frame features used in the LCQA algorithm.

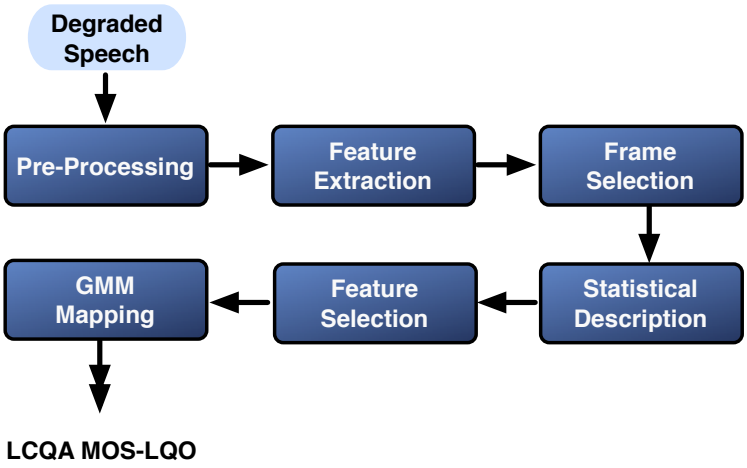

Figure 3.3.3: The overall structure of the LCQA algorithm.

3.4 Per-utterance Quality

In this section, the problem of estimating the quality of an utterance of speech is considered. The assumptions are that the speech signal can easily be decomposed into speech utterances of a short length (here signals with a duration between 3 and 8 seconds are considered). An utterance is defined as a segment of speech for which the measure of interest is constant. In Section 3.5, the problem of estimating the time-varying quality of speech is considered with speech material of 60 seconds duration and block-varying SNR.

Sections 3.4.1 and 3.4.2 present two extensions of the LCQA algorithm. This is followed by a description of the NISQ algorithm, which is based on the NISA framework. This is followed by a validation of the performance of the PESQ algorithm on the C-Qual database in Section 3.4.4. The databases and evaluation metrics are presented in Sections 3.4.5 and 3.4.6 respectively. This subsection is concluded with results in Section 3.4.7.

3.4.1 LCQA2

This is a further development of the LCQA algorithm that utilizes the same feature extraction framework (modeling per-frame features with their statistical properties), the LPC based feature set and the GMM modeling. The novel extensions are 4 additional per-frame features, the use of a noise robust pitch estimation algorithm, an external VAD[4] and a two-step feature selection and projection technique. The input signal is divided into frames of 20 ms duration without overlap and windowed by a Hanning window. The energy per frame is normalized to make the following feature extraction gain independent. A total of 15 features are extracted for each frame, referred to as per-frame features. The mean, variance, skewness and kurtosis of each per-frame feature is used to characterize the input signal properties, referred to as global features. A two-step dimensionality reduction using the raw feature correlations and principal component analysis (PCA) is applied to the global feature set to select the optimum features. A GMM is trained on the joint density of the optimum features and the MOS as in LCQA [53].

Features

The LCQA2 algorithm uses the entire set of 11 per-frame features from LCQA, with replacement of the original pitch estimation algorithm by the PEFAC method. The PEFAC algorithm was

[4] as described in Chapter 2

shown to have a high robustness to additive noise at low SNRs in Chapter 2. The zero crossing rate and the iSNR features and their rate of change are included as additional per-frame features. The per-frame feature vector has 15 features per frame of the signal (Table 3.5) and these are characterized by the mean, variance, skewness and kurtosis of the per-frame features as in LCQA. The resulting feature vector per signal has a dimension of 60[5].

Dimensionality reduction

The LCQA2 algorithm performs a two step dimensionality reduction scheme based on a feature subset selection followed by a feature projection step (on the training data). The first stage is a feature subset selection, which is achieved through a correlation analysis of the features. It is desirable to retain only those features that have a high correlation with the quality score and at the same time, are uncorrelated with other features. The correlation coefficient based measure for feature m is obtained as :

$$F_{cor}(m) = \frac{r_p(\Phi_m, \Theta)}{\sum_{j \neq m} r_p(\Phi_m, \Phi_j)}, \tag{3.4.1}$$

where $r_p(\Phi_m, \Theta)$ is the Pearson correlation coefficient between the vector containing all examples of the m^{th} feature and the vector of labels for each signal (Θ), defined as

$$r_p(X,Y) = \frac{\sum_{n=1}^{N}(X_n - \mu_X)(Y_n - \mu_Y)}{\sum_{n=1}^{N}(X_n - \mu_X)^2 \sum_{n=1}^{N}(Y_n - \mu_Y)^2}, \tag{3.4.2}$$

where μ_X is the mean of the vector quantity X and N is the dimension of X and Y. A set of $\hat{N}$ features are selected from the N global features, based on the rankings obtained by the correlation measure[6]. This is followed by performing a PCA based feature projection by retaining Q linear combinations of the M optimal features, such that the total variance accounted for by the projected features is greater than 95%[7]. The resulting Q features are used to train an M mixture GMM according to the original LCQA framework [53].

[5] N=60.

[6] this number is optimized experimentally from the training data.

[7] i.e. Q is chosen as the minimum number of Eigenvectors that account for 95% of the variance in the Eigenvalues.

Feature description	Feature	Rate of change of feature
Spectral flatness	ϕ_1	ϕ_9
Spectral dynamics	ϕ_2	-
Spectral centroid	ϕ_3	ϕ_{10}
Zero crossing rate	ϕ_4	ϕ_{11}
Excitation variance	ϕ_5	ϕ_{12}
Speech variance	ϕ_6	ϕ_{13}
Pitch period	ϕ_7	ϕ_{14}
iSNR	ϕ_8	ϕ_{15}

Table 3.5: The 15 per-frame features used in the LCQA2 algorithm.

3.4.2 LCQA-M

The second development of the LCQA method uses the mel-frequency cepstral coefficients (MFCC) derived from the short-time FFT as the main features and follows the framework described in Section 3.4.1. The MFCCs represent the perceptually relevant aspects of the short-term speech spectrum and have been shown to give a good performance for speech recognition [27, 90] and speaker verification and recognition [17, 121]. In the LCQA-M method, the LPC derived features are replaced with the MFCCs as well as their velocity and acceleration features. The zero crossing rate, pitch frequency and iSNR and their rate of change over all frames are also included as per-frame features. The resulting per-frame feature vector contains 42 features per frame as outlined in Table 3.6. The mean, variance, skewness and kurtosis of the per-frame features are used to characterize the per-frame features, resulting in 168 global features. A two step dimensionality reduction is performed as in the LCQA2 method, followed by a GMM mapping.

Feature description	Feature	Rate of change of feature
MFCCs	$\phi_{1:12}$	-
MFCC - Velocity	$\phi_{13:24}$	-
MFCC - Acceleration	$\phi_{25:36}$	-
Zero crossing rate	ϕ_{37}	ϕ_{40}
Pitch frequency	ϕ_{38}	ϕ_{41}
iSNR	ϕ_{39}	ϕ_{42}

Table 3.6: The 168 per-frame features used in the LCQA-M algorithm.

3.4.3 NISQ

The NISQ algorithm is the application of the NISA framework for speech quality estimation. The 25 per-frame features used in the NISQ algorithm are shown in Table 3.7. The LPC based features ($\phi_{1,2,3,5}$) have been used for non-intrusive speech quality assessment in the LCQA algorithm [53] and are also included in the NISQ algorithm. The zero crossing rate has been successfully used as a feature for voiced-unvoiced speech and silence classification [9] and is also expected to be a useful feature for speech quality assessment. The iSNR, Hilbert envelope and PLD based features were presented in Section 2.4 and are expected to be particularly useful for modeling the effects of additive noise on speech quality. The 25 per-frame features are characterized by their mean, variance, skewness and kurtosis, resulting in 100 global features. Additionally, 16 features characterizing the the long-term spectral deviation are calculated, resulting in 116 global features, which are used to train a CART regression tree.

Feature description	Feature	Rate of change of feature
Spectral flatness (LPC)	ϕ_1	ϕ_{14}
Spectral dynamics (LPC)	ϕ_2	-
Spectral centroid (LPC)	ϕ_3	ϕ_{15}
Zero crossing rate	ϕ_4	ϕ_{16}
Excitation variance	ϕ_5	ϕ_{17}
Speech variance	ϕ_6	ϕ_{18}
Pitch period	ϕ_7	ϕ_{19}
iSNR	ϕ_8	ϕ_{20}
Hilbert envelope variance	ϕ_9	ϕ_{21}
Hilbert enveloped dynamic range	ϕ_{10}	ϕ_{22}
Spectral flatness (PLD)	ϕ_{11}	ϕ_{23}
Spectral dynamics (PLD)	ϕ_{12}	ϕ_{24}
Spectral centroid (PLD)	ϕ_{13}	ϕ_{25}

Table 3.7: The 25 per-frame features used in the NISQ algorithm.

The long term features are calculated as the deviation of the long term magnitude spectrum of the current signal from LTASS. The resulting residual magnitude spectrum P_{LTLD} is then mapped into 8 bins (equal bandwidth, 50% overlap), each with a bandwidth of 500 Hz. The energy in each bin as a percentage of the total energy is then computed and forms the long term features in NISQ.

$$\phi_j = \frac{\sum_{g \in w} P_{LTLD}(g)}{\sum_{k=1}^{K} P_{LTLD}(k)},$$

where $j = [26, ..., 41]$ and w is a 500 Hz window centered on the frame of interest and the numerator is the energy of the current frame and the numerator is the total energy in the

residual spectrum.

3.4.4 Validation of PESQ

In order to use PESQ as a ground truth labeling algorithm, a validation of PESQ was performed on the C-Qual database using the C implementation from the ITU-T (validated using the PESQ validation data fromm ITU-T). The levels of degradation in C-Qual were selected to cover the range of PESQ scores from approximately 0.8 to 4.5, for each degradation type. The estimated PESQ scores are compared with subjective MOS to validate the performance of PESQ in the context of law enforcement. The Spearman rank correlation coefficient (r_s) [104] between the objective quality scores and subjective MOS obtained from the listening experiment was used to validate the performance of wide-band PESQ [85, 88] algorithm. The motivation for using r_s is that this metric is not effected by the monotonic mapping functions often employed in optimizing the PESQ scores to the MOS-LQS obtained from listening tests [86]. The objective for evaluation here is to test how well the raw PESQ scores correlate with MOS-LQS, independent of a mapping function.

Additionally, the Root Mean Square Error (RMSE) was computed for the mapping function proposed in [86] for mapping the PESQ scores to MOS-LQO. Although PESQ has been shown to attain a high correlation with subjective MOS for telecommunications distortions, it performs poorly on the degradation conditions present in the C-Qual database, as supported by a low overall correlation coefficient of 0.57 and an RMSE of 1.19 MOS. If the database conditions are split into additive noise (conditions 1 to 21) and non-linear degradations (conditions 22 to 44), then one can get a better idea of where PESQ fails. As shown in Table 3.8, a high correlation of 0.93 is obtained for the additive noise conditions (between PESQ and MOS-LQS). However, the RMSE for the mapping function is still higher than 1 MOS unit[8] making it unsuitable for use in this application. It can be seen that PESQ is poor at predicting the effects of the other distortion conditions such as peak clipping, reverberation and coloration, with a correlation of 0.16 (between PESQ and MOS-LQS). This limits the use of wide-band PESQ to additive noise conditions and the original mapping function cannot be used for the levels of degradation present in C-Qual. As a result, the extended-database presented in Section 3.4.5 is limited to additive noise and telecommunication channel based degradations.

	All conditions		Additive noise		Other conditions	
	PESQ	PESQ MOS-LQO	PESQ	PESQ MOS-LQO	PESQ	PESQ MOS
r_s	0.57	0.57	0.93	0.93	0.16	0.16
RMSE	-	1.19	-	1.15	-	1.23

Table 3.8: The performance of Wide-band PESQ for different groups of conditions from the C-Qual database.

[8] this could be the difference between good and poor quality speech if the ground truth quality score was in the MOS=3 region.

3.4.5 Databases

This section presents the speech quality databases used for evaluating the non-intrusive algorithms. The training and validation database (referred here as the TN database) is based on speech from the TIMIT database [48], which is a phonetically diverse database with speech from 623 American English speakers representing eight dialect regions across the USA. The speech material consists of 10 utterances by each of the speakers, of which 8 are unique and 2 are common to all speakers. In the TN database, only the distinct utterances are used for all the speakers. An extensive additive noise database is then created by adding 15 noises from the from the NATO noise database [167] at SNR's in the range -24 to 30 dB in 3 dB steps, with speech level calculated using the P.56 method [82]. Additionally, for the purpose of evaluating the quality effects of real telecommunication channels, the CTIMIT [22] and NTIMIT [93] databases are also included to represent realistic telecommunications degradations. The NTIMIT contains speech from the original TIMIT database transmitted through the telephone network and recorded at the listener end. Similarly, the CTIMIT comprises of TIMIT utterances transmitted through the cellular network. The resulting TN database is composed of 285 additive noise conditions[9] for each of speaker and the union of the CTIMIT and NTIMIT databases.

The C-Qual database [146] includes the types and levels of degradations commonly found in law enforcement applications, with speech material composed of the English partition of the ITU-T P.23 database [77]. The non-intrusive experiments reported in this thesis make use the additive noise partition of the C-Qual database, comprising of car, babble and hum noise representing 21 conditions for each speaker (of which there are 4). This is used only as a generalization test database as it contains different speech and noise material. All databases were down-sampled to 8 kHz to represent narrowband speech transmission.

3.4.5.1 Quality labeling

The labelling of the databases for the purpose of speech quality testing is carried out with the PESQ algorithm. In the case of law enforcement relevant degradations, PESQ works well for the additive noise conditions [146], achieving a Spearman correlation coefficient of 0.93 with subjective quality scores in this scenario. Since the TN database only contains additive noise and telecommunication type distortions (distortions primarily due to transmission channel), the PESQ score is expected to be well correlated with speech quality and is therefore used as the ground truth for the non-intrusive quality experiments.

[9] due to 19 SNR's $\times$ 15 noises per speaker.

3.4.5.2 Training

The original TIMIT database is partitioned into a training and test partition, which is maintained throughout the degradation processing explained in Section 3.4.5. The training partition contains 455 speakers and the remaining 168 speakers in the test partition. There is no overlap of speech material in the test and training set (different text and speakers). A similar partitioning is maintained in the TN database, with the training partition consisting of 168 speakers randomly selected from the 455 speakers in the training partition. All data-driven algorithms are trained on the TN training partition and also tested on the entire test partition. Additionally, for the additive noise partition, each noise file is also split into a training and test partition to ensure the same noise source is not used in the training and test partitions. The resulting TN database consists of more than 45 hours of speech in the training and test partitions.

3.4.6 Evaluation Metrics

This section defines the metrics used for measuring the performance of algorithms for speech quality assessment. As the ground truth label for the test and training databases outlined in Section 3.4.5 is carried out using the intrusive PESQ algorithm, the task for the comparative evaluation is of non-intrusive PESQ estimation. The metrics used in the evaluation are outlined as follows.

Spearman Correlation Coefficient (SCC)

The Spearman rank correlation coefficient is a non-parametric measure that describes the monotonic relationship between two ranked variables [104] and is calculated as the Pearson correlation coefficient between the ranked variables as follows

$$r_s(X,Y) = r_p(\check{X},\check{Y}),$$

where $\check{X}$ and $\check{Y}$ are the ranks of X and Y and r_p is the Pearson correlation coefficient. The quality scores obtained from subjective listening experiments are subject to MOS scale variations due to a number of reasons, for example the quality range included in an experiment and the subject's cultural background. To compensate for such systematic variations a mapping function between the objective quality scores and MOS-LQS is often used [87]. The use of the rank correlation coefficient makes the analysis independent of the mapping function and allows a comparison of

the rank similarity of the algorithms to be performed.

Root Mean Square Error (RMSE)

The root mean square error between the estimated and true scores is calculated as a measure of the estimation accuracy of each algorithm as follows

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(\varepsilon_n)^2}{N}},$$

where ε_n is the residual error defined as

$$\varepsilon_n = \theta_n - \hat{\theta}_n, \tag{3.4.3}$$

where θ_n and $\hat{\theta}_n$ are the ground truth and estimated PESQ scores for the nth signal respectively.

Bin Error

The bin error evaluates the absolute mean residual error in PESQ bins of size 0.25 for the PESQ prediction performance. This metric shows the percentage of signals that lie in each PESQ bin and provides a view of the frequency of errors of different magnitudes, similar to the case of MOS prediction [85, 87].

Two Class Classification (TCC)

This measure investigates the hit rate (HR) achieved by splitting the ground truth scores into two classes (set to the mid point of the relevant scale). The motivation for this metric is to evaluate the algorithms in terms of a good quality or bad quality criteria, where an acceptance threshold is set to be the midpoint of the PESQ scale. This metric is particularly useful for quality assessment where the user may want to detect the degradation of the system below or above an acceptance threshold, which in this case is set to a PESQ score of 2.

3.4.7 Results

This section presents the results for non-intrusive PESQ estimation for the P.563, LCQA, LCQA2, LCQA-M and NISQ algorithms on the databases described in Section 3.4.5. The results for the P.563 algorithm reported here are based on the ITU-T's published C implementation. The LCQA, LCQA2, LCQA-M and NISQ algorithms have been implemented in Matlab by the author. Table 3.9 shows the results for PESQ estimation on the test partition of the TN database. The best performance is achieved by the NISQ algorithm on all the test metrics, with an SCC of 0.90 and an RMSE of 0.4. Moreover, 95% the estimation errors for the NISQ algorithm are less than 0.75 PESQ score. The LCQA-M and LCQA2 algorithms have a similar performance to the LCQA algorithm, with a small improvement in SCC and RMSE with all algorithms outperforming the ITU standard non-intrusive P.563 method. The generalization performance of the algorithms is ascertained by testing on the additive noise partition of the C-Qual database labeled with PESQ and the results for this are shown in Table 3.10. It can be seen that NISQ outperforms the other methods with a marginally lower SCC and higher RMSE than for the TN database, this is attributed to the difference in the speech and noise material in C-Qual. The LCQA2 and LCQA-M algorithms perform better than the LCQA algorithm in the generalization test and this may be due to the additional features in these methods being more discriminative in the task.

The best 10 features for the LCQA2, LCQA-M and NISQ algorithms is presented in Table 3.11, which allows a comparison of the importance of the different features for non-intrusive PESQ estimation. The feature selection for LCQA2 and LCQA-M algorithms is independent of the machine learning algorithm as the measure is a normalized correlation between the features and the PESQ score. It can be seen that the most important feature in the LCQA2 method is the mean of the spectral dynamics of the LPC magnitude spectrum and the iSNR and pitch are important features. In the LCQA-M method, the mean pitch and the variance of the delta and delta-delta features are important. This supports the view that the dynamics of the signal is an important feature, as in speech coding where the dynamics of the envelope is more perceptually audible than spectral distortion [101] and similarly for speech enhancement [131]. In the case of NISQ, the SNR, pitch and the LTASS deviation based features are important for PESQ estimation.

The NISQ algorithm uses 25 features for binary tree regression model with the dynamic range and variance of the Hilbert envelope also being useful features. The zero crossing rate and LPC residual are not used. The LCQA+ model retains 16 features after feature selection and 11 linear combinations are used after feature projection. The LCQA-M model retains 12 features after feature selection and 7 linear combinations are used after feature projection and

the LCQA algorithm selects 12 features after feature selection. The number of features and linear combinations are obtained experimentally by re-substitution estimation on the training data.

			Bin Error				TCC
Algorithm	**SCC**	**RMSE**	**<0.25**	**<0.5**	**<0.75**	**<1.0**	**HR**
NISQ	0.90	0.40	58.0	86.8	95.3	97.9	92.8
LCQA-M	0.86	0.55	37.2	65.8	83.4	92.9	90.7
LCQA2	0.85	0.58	38.7	67.9	86.7	93.5	90.5
LCQA	0.83	0.57	40.0	68.4	84.8	93.4	90.9
P563	0.81	0.94	21.8	43.7	61.6	74.7	88.9

Table 3.9: Non-intrusive PESQ estimation performance on the test partition of the TN database.

			Bin Error				TCC
Algorithm	**SCC**	**RMSE**	**<0.25**	**<0.5**	**<0.75**	**<1.0**	**HR**
NISQ	0.88	0.43	59.5	83.3	98.8	100	92.9
LCQA-M	0.85	0.61	29.8	53.6	73.8	90.5	77.4
LCQA2	0.80	0.64	29.8	60.7	76.2	85.3	69.1
LCQA	0.74	0.66	23.8	57.1	75.0	84.5	64.3
P563	0.60	0.96	22.7	37.5	51.1	69.9	88.1

Table 3.10: Non-intrusive PESQ estimation performance on the additive noise partition of the C-QUAL database, representing a generalization test.

Rank	**LCQA2**	**LCQA-M**	**NISQ**
1	$\mu(\phi_2)$	$\mu(\phi_{38})$	$\mu(\phi_8)$
2	$\mu(\phi_8)$	$\sigma(\phi_1)$	$\sigma(\phi_{23})$
3	$\mu(\phi_6)$	$\sigma(\phi_{27})$	$\mu(\phi_{11})$
4	$\sigma(\phi_{15})$	$\sigma(\phi_{15})$	$\mu(\phi_7)$
5	$\kappa(\phi_8)$	$\sigma(\phi_{18})$	$\sigma(\phi_{13})$
6	$\gamma(\phi_7)$	$\sigma(\phi_{30})$	$\mu(\phi_2)$
7	$\sigma(\phi_4)$	$\sigma(\phi_{17})$	$\kappa(\phi_{23})$
8	$\sigma(\phi_{10})$	$\sigma(\phi_6)$	ϕ_{27}
9	$\gamma(\phi_2)$	$\sigma(\phi_{29})$	$\mu(\phi_6)$
10	$\kappa(\phi_{15})$	$\sigma(\phi_{31})$	ϕ_{31}

Table 3.11: The 10 best ranked features for non-intrusive PESQ estimation based on the training partition of the TN database.

3.5 Time varying assessment

In this section, the time-varying assessment of speech quality is presented without explicit utterance splitting for long speech recordings. The research questions addressed here are threefold. First to measure the performance of speech quality estimation in short time blocks as a function of block-size; second to evaluate how features derived from Mel-Frequency Cepstrum (MFCC) and Linear Predictor Coefficients (LPC) compare in terms of speech quality estimation as a function of block-size; finally to evaluate how errors in objective estimation are distributed over the entire range of MOS. Moreover, the speech material used in the evaluation consists of long recordings with a block-varying SNR.

3.5.1 Algorithms

The algorithms evaluated are the same as in the previous section on per-utterance, non-intrusive PESQ prediction. The only change in this section is that an explicit VAD is not used but instead, any frames where there is no speech present are assigned a MOS of 1. This follows from the observation that in a speech pause region, there is likely to be only background noise present and thus should be assigned the lowest quality rating. This framework simplifies the system evaluation as estimating quality on the MOS scale of 1 to 5, without explicitly identifying speech pause regions, since the aim is to perform the estimation of long speech recordings.

3.5.2 Methodology

The evaluation of the objective methods is carried out using a block-varying extension of the additive noise conditions from the C-Qual database [146], referred as the TVC-Qual database. The TVC-Qual database is split into a training and test partition, with speech from 2 speakers each partition. A 50% cross-validation is adopted for the testing, whereby the training set uses speech from male and female speaker A and test set uses the male and female speaker B. Then the experiment is repeated with A and B swapped and the results combined. The test set always excludes data from the training set. The algorithms are tested for block sizes in the range 0.5 to 8.0 seconds in 0.5 second steps and for each block size, a number of individual blocks are estimated (i.e. for a block size of 1 s there will be 60 blocks) for each speech file. The training of the data-driven algorithms is performed separately for each block size.

3.5.2.1 Database

A block-varying extension of the C-Qual database is achieved by concatenating sentences (each of 4 second duration, with an active speech length of 3.5 s) from the same speaker degraded by the same noise type, using a 10 ms crossfade in the speech pause regions at the beginning and end of the sentences. The resulting files have a duration of 60 s and a block-varying[10] SNR. The database contains 84 minutes of speech corresponding to 2 male and 2 female speakers, representing 21 additive noise conditions, including car and babble noise at 7 SNRs per noise type. Each file contains nearly 60 seconds of speech from a single speaker and a singe noise type, with a random fluctuation of SNR. A total of seven SNR profiles are included for each noise type and speaker. Figure 3.5.1 shows 3 example SNR profiles for a female speaker, with per-condition MOS from the C-Qual database. The quality score for a block is calculated as the average of the subjective MOS for the segments being concatenated, resulting in a piece-wise constant, block varying MOS profile.

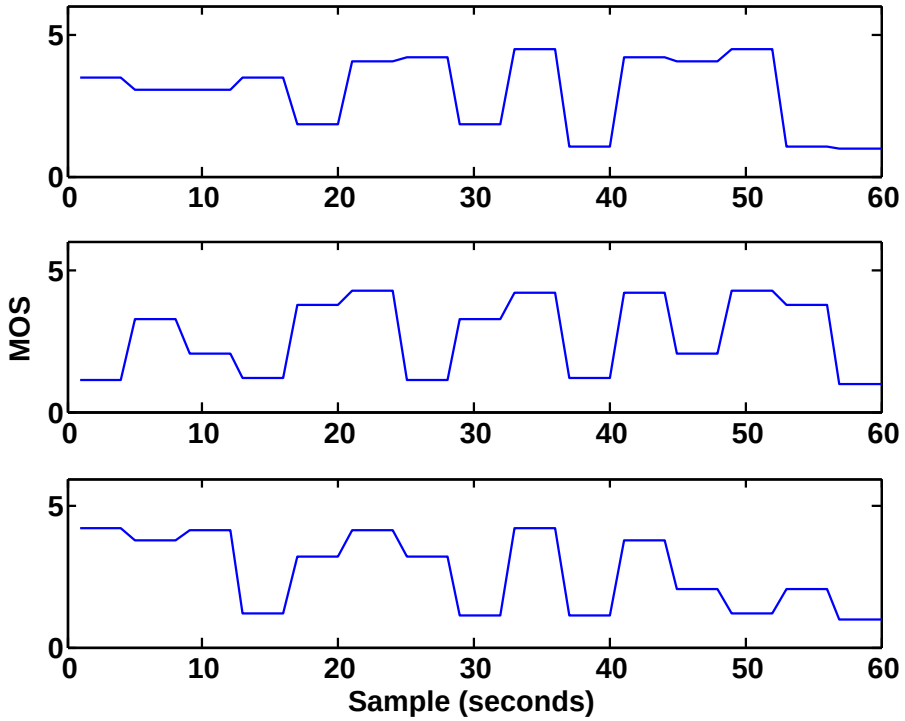

Figure 3.5.1: The block-varying MOS profiles for 3 example SNR profiles. The x-axis shows the block samples in seconds, each of which is approximately 4 seconds in length.

[10]The SNR is varying in 4 second blocks, rather than smoothly time-varying.

3.5.2.2 Evaluation metrics

The overall performance is measured using the average Pearson correlation coefficient between the MOS and the estimated MOS for all files evaluated with a particular block size. The average correlation over all files is used as a figure of merit, defined as:

$$PCC = \frac{1}{N_j}\sum_j r_p^{(j)}, \tag{3.5.1}$$

where N_j is the number of files to be tested (for a particular block size) and $r_p^{(j)}$ is the Pearson correlation coefficient between the estimated and ground truth MOS for every block in file j, defined as:

$$r_p^{(j)} = \frac{\sum_n(\hat{\theta}_n - \mu_{\hat{\theta}})(\theta_n - \mu_\theta)}{\sqrt{\sum_n(\hat{\theta}_n - \mu_{\hat{\theta}})^2 \sum_n(\theta_n - \mu_\theta)^2}}, \tag{3.5.2}$$

where θ_n is the MOS and $\hat{\theta}_n$ is the estimated MOS for block n.

In addition to the average correlation between the MOS and estimated MOS, the RMSE is also used as a measure of estimation accuracy. It is also advantageous to analyze the distribution of the errors over the MOS range of 1 to 5. In many situations, errors in the range of quality scores of 3 to 5 are less significant than errors in the range of 1 to 3. The RMSE-Bin is a measure of the root-mean-square error (RMSE) between the subjective and objective scores, calculated over 2 MOS bins ($[1 \leqslant \theta < 3]$, $[3 \leqslant \theta \leqslant 5]$) representing the estimation error in poor and good quality speech respectively.

3.5.3 Results

This section presents the results for non-intrusive estimation of speech quality (MOS-LQS) for block-varying degradations of long speech signals (60 s). The performance of the algorithms for non-intrusive quality estimation in terms of the PCC metric is presented in Fig. 3.5.2 which shows that the NISQ algorithm outperforms all other methods tested in the block-varying MOS estimation task. Moreover, the best performance is achieved for a block size of 1 second, with a correlation of 0.98 and RMSE of 0.2 MOS. A similar conclusion can be drawn from Fig. 3.5.3 which shows the estimation accuracy in terms of the RMSE metric, where the NISQ algorithm has an RMSE of less than 0.5 MOS for all block sizes tested. Moreover, the RMSE accuracy of NISQ is very similar in the low MOS bins (Fig. 3.5.4) and the high MOS bins (Fig. 3.5.5).

The LCQA2 and LCQA-M algorithms have a similar performance with a higher PCC (and

lower RMSE) for larger block sizes, with an RMSE lower than 0.5 MOS for block sizes greater than 3.5 seconds. Moreover, these developments of the LCQA perform better than the baseline LCQA method for block sizes greater than 1.5 s. The LCQA algorithm has a good performance in the 0.5 to 2.0 s region, with a peak PCC of 0.80 for a block size of 1.5 s and a corresponding RMSE of 0.68. This is in contrast with the P.563 performance, which is particularly low for block sizes less than 1.5 s with PCC of 0.42 and RMSE of 2.2 for a block size of 0.5 s. The main region of degradation for the P563 algorithm is in the lower MOS bins (MOS in the 1 to 3 region) with an RMSE of 2.6 as shown in Fig. 3.5.4.

Table 3.12 shows the best feature for the LCQA2, LCQA-M and NISQ algorithms for different block sizes. The LPC derived spectral dynamics is the most important feature in the LCQA2 and NISQ algorithms, with the mean of this feature good for smaller block sizes (up to 3.0 to 4.0 s) and the variance of this feature important for the larger block sizes. Similarly for the LCQA-M algorithm, the velocity and acceleration of the MFCC's are the important features. This is in line with the per-utterance case, where the dynamics of the signal are important features.

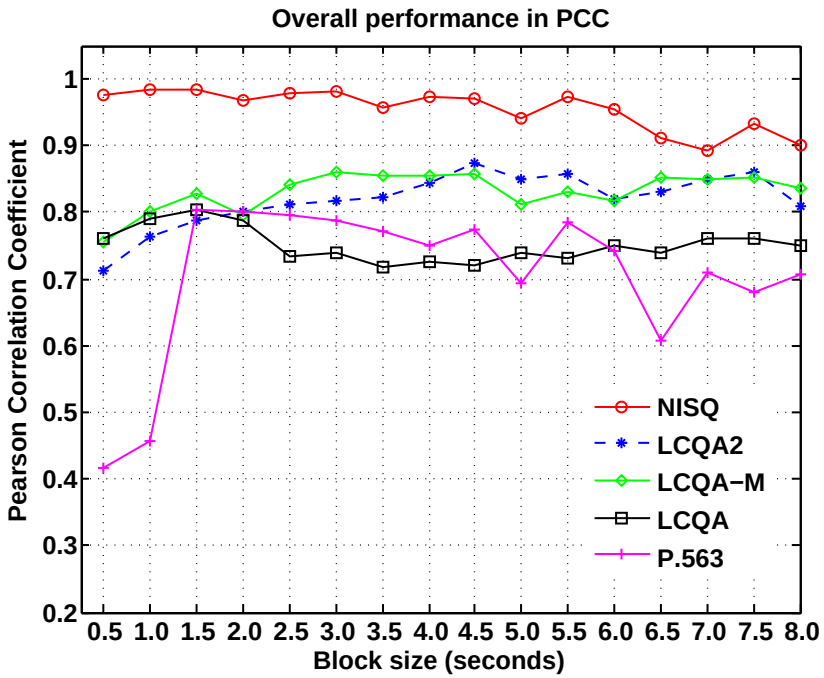

Figure 3.5.2: Performance of non-intrusive estimation of block-varying MOS using the PCC metric.

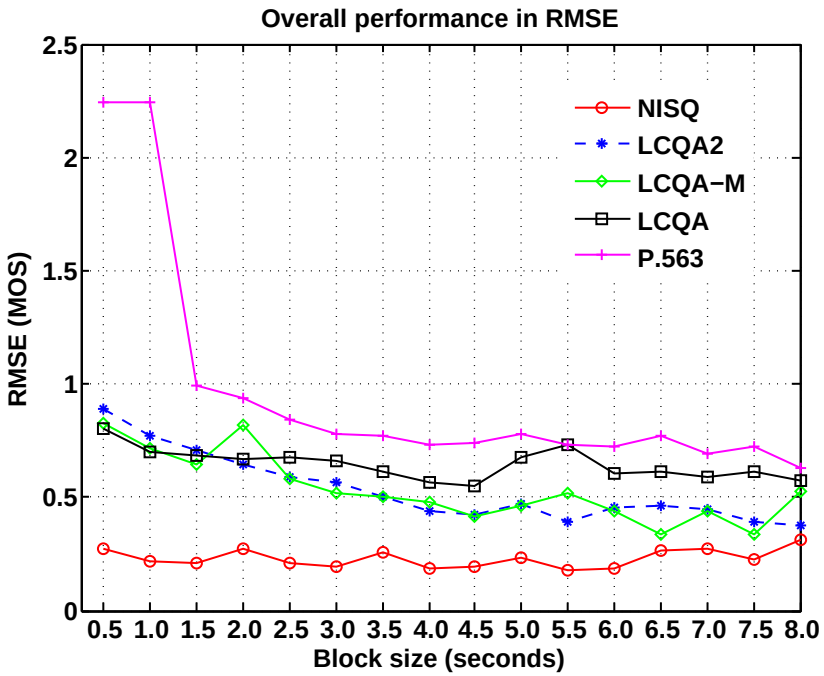

Figure 3.5.3: Performance of non-intrusive accuracy of block-varying MOS using the RMSE metric

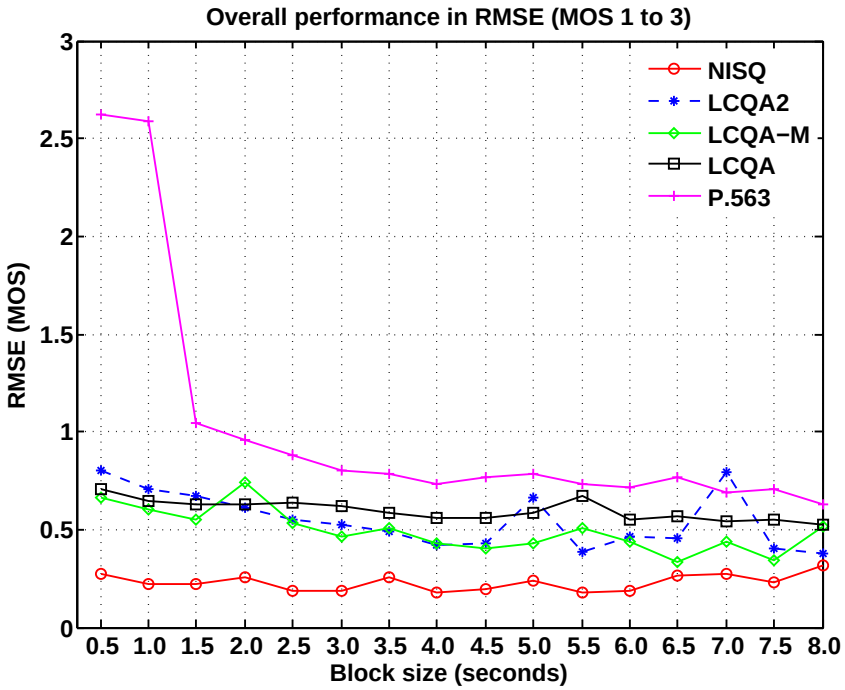

Figure 3.5.4: Estimation accuracy (RMSE) in MOS bins 1 to 3 for block-varying C-Qual.

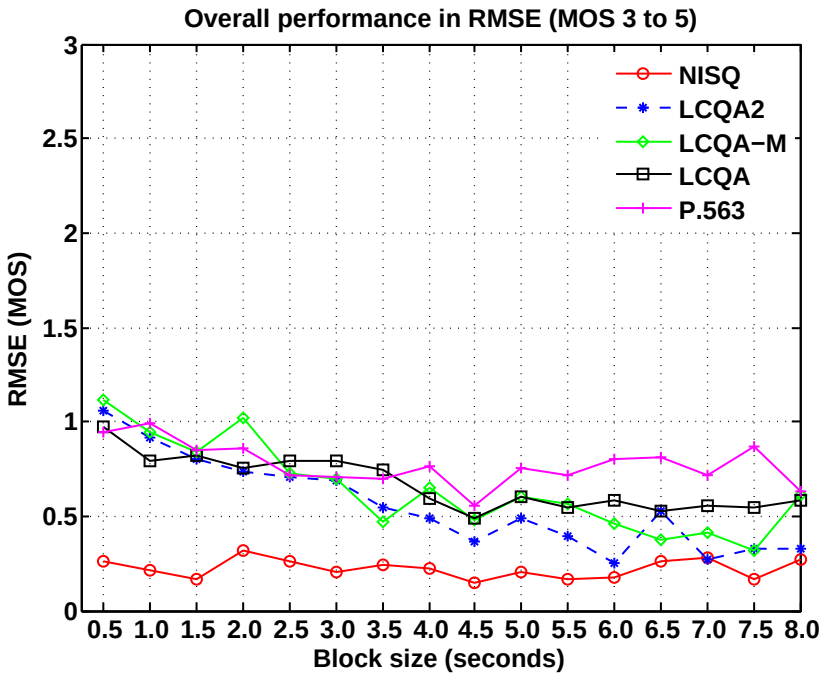

Figure 3.5.5: Estimation accuracy (RMSE) in MOS bins 3 to 5 for block-varying C-Qual.

Block (s)	LCQA2	LCQA-M	NISQ
0.5	$\mu(\phi_7)$	$\sigma(\phi_{19})$	$\mu(\phi_2)$
1.0	$\mu(\phi_2)$	$\sigma(\phi_{17})$	$\mu(\phi_2)$
1.5	$\mu(\phi_2)$	$\sigma(\phi_{19})$	$\mu(\phi_2)$
2.0	$\mu(\phi_2)$	$\sigma(\phi_{16})$	$\mu(\phi_2)$
2.5	$\mu(\phi_2)$	$\sigma(\phi_{17})$	$\mu(\phi_2)$
3.0	$\mu(\phi_2)$	$\sigma(\phi_{17})$	$\mu(\phi_2)$
3.5	$\sigma(\phi_2)$	$\sigma(\phi_{19})$	$\mu(\phi_2)$
4.0	$\sigma(\phi_2)$	$\sigma(\phi_{19})$	$\mu(\phi_2)$
4.5	$\sigma(\phi_2)$	$\mu\phi_2)$	$\sigma(\phi_2)$
5.0	$\sigma(\phi_2)$	$\gamma(\phi_{36})$	$\sigma(\phi_2)$
5.5	$\sigma(\phi_2)$	$\gamma(\phi_{34})$	$\sigma(\phi_{25})$
6.0	$\sigma(\phi_2)$	$\gamma(\phi_{34})$	$\sigma(\phi_{25})$
6.5	$\sigma(\phi_2)$	$\sigma(\phi_{23})$	$\sigma(\phi_{25})$
7.0	$\sigma(\phi_2)$	$\sigma(\phi_{23})$	$\sigma(\phi_2)$
7.5	$\sigma(\phi_2)$	$\mu(\phi_{10})$	$\sigma(\phi_2)$
8.0	$\sigma(\phi_2)$	$\mu(\phi_{36})$	$\sigma(\phi_2)$

Table 3.12: The best feature at each block-size for the LCQA2, LCQA-M and NISQ algorithms.

3.6 Summary

The problem of non-intrusive speech quality estimation in the law enforcement context was presented in this chapter. Given the requirement for large quantities of training data for developing data-driven algorithms, the intrusive PESQ algorithm was used as a ground truth labeling method. The performance of the PESQ algorithm was validated on the C-Qual database, where it was shown that PESQ can be applied to the additive noise conditions, for which a Spearman correlation coefficient of 0.93 was obtained between the raw PESQ scores and MOS-LQS. The performance of PESQ for the other distortions in C-Qual was very poor (correlation of 0.16 with MOS-LQS) and thus the extended-database did not contain these distortions. This motivated the development of a training and validation database based on speech from 336 speakers from the TIMIT database and 15 noises from the Nato database. The C-TIMIT and N-TIMIT databases were also included to form a large database with high levels of additive noise and degradations due to realistic communication channels. The evaluation of speech quality was presented under two scenarios, one being the traditional per-utterance case and the other being the non-intrusive time-varying quality estimation.

The LCQA algorithm was improved by adding new features and a two step dimensionality reduction scheme resulting in the LCQA2 and LCQA-M algorithms, which outperform the baseline LCQA and ITU standard P.563 methods. A novel algorithm based on the NISA framework, referred as NISQ was also proposed and shown to give the highest performance in non-intrusive quality estimation with a correlation of 0.90 with PESQ. Similar conclusions were drawn from the tim-varying experiments on the TVC-Qual database, where the NISQ algorithm was shown to have an RME lower than 0.5 MOS for block sizes in the 0.5 to 8.0 second range. The features capturing the dynamics of the spectrum were found to be important for speech quality estimation, together with the iSNR, pitch and LTASS deviation features.

Chapter 4

Speech Intelligibility Assessment

THIS chapter presents the problem of non-intrusive speech intelligibility assessment, beginning with an introduction of the concept of speech intelligibility and then a review of subjective and objective techniques for obtaining intelligibly scores. The problem of non-intrusive intelligibility estimation of noisy speech is presented in Section 4.3 followed by an initial study on non-intrusive assessment of noise suppressed speech in Section 4.4. The research presented in this chapter relates in part to the following publications [145, 149].

4.1 Introduction

Speech intelligibility is a measure of the proportion of a speech signal correctly recognized by a listener. In contrast to speech quality, the intelligibility of a speech signal can be measured and is not only a perceived construct. In certain speech assessment techniques, speech intelligibility is considered to be an aspect of speech quality, as in the diagnostic acceptability measure [169]. It is an important quantifier for applications such as telecommunications, where the quality of a channel may be evaluated in terms of its effect on speech intelligibility [154], as a performance metric for hearing aids [166], for determining the impact of an acoustic space on speech [72] and for intelligence gathering in law enforcement applications [125]. Moreover, the development of intelligibility assessment algorithms for noise suppressed speech (such as [157]) have made it possible to quantify objectively the effects of speech enhancement on intelligibly. This is a particularly important issue in law-enforcement as recent studies on the effects of noise suppression on speech intelligibility indicate the many such techniques have an adverse effect

on speech intelligibility [74, 66].

The remainder of this chapter is organized as follows. A review of current methods for speech intelligibility assessment and methods for assessing the effects of noise suppression on speech intelligibility is presented in Section 4.2. This is followed by an investigation of non-intrusive speech intelligibility assessment in Section 4.3, followed by a study on estimating the effects of noise suppression on speech intelligibility in Section 4.4.

4.2 Review

4.2.1 Intelligibility assessment

A number of methods have been proposed in the literature for obtaining speech intelligibility scores and these may be classified as either subject based or objective. Subjective speech intelligibility scores are obtained through listening experiments where subjects listen to speech samples and their performance in a particular linguistic task is measured. The linguistic task may be to recognize nonsense syllables, isolated words or a number of key-words in a sentence. In phonetic intelligibility tests, the task is to recognize syllables and examples include [46] and [118]. In word intelligibility tests, the task is to recognize phonetically balanced words such as in [32, 170], which is closer to the actual scenario of typical speech communication. The task in sentence intelligibility tests is to identify key-words in a sentence, which is yet more realistic than the isolated word intelligibility task and examples include the Speech Perception in Noise (SPIN) test [96] and the Hearing in Noise test (HINT) [126]. It is necessary to perform subject based experiments on many subjects in order to get a reliable estimate of the intelligibility scores, which makes the task of obtaining subjective intelligibility scores expensive and time consuming.

The characteristic shape of the relationship between noise level and speech intelligibility takes the form of a sigmoid psychometric function (PF), commonly characterized by a slope, guess rate and lapse rate [50, 100]. A common technique in intelligibility testing is to measure the SNR corresponding to 50% intelligibility (a single point on the PF), referred to as the Speech Reception Threshold (SRT) [127] and the Bayesian Adaptive Speech Intelligibility Estimation (BASIE) procedure that allows rapid intelligibility estimation of speech in noise using digit-triplets has been recently proposed [50]. However, even such methods require a number of subjects to perform the testing and are time consuming and expensive for large quantities of speech material.

Objective intelligibility assessment methods operate without the need for human subjects and can provide rapid intelligibility scores. The current focus in the literature is on intrusive methods and as far as the author is aware, at the time of writing of this thesis there is only one no non-intrusive intelligibility assessment technique, which was proposed by the author in [145]. The intrusive objective methods assume that a clean signal (with 100% intelligibility) is available in addition to the degraded signal and the focus of the development has been on modeling the effects of additive noise and reverberation. One of the earliest intrusive, objective intelligibility technique was proposed by French and Steinberg [47] as the Articulation index (AI), which was

further refined with improved methods for calculating the AI proposed by Kryter in 1962 [103] and finally led to an ANSI standard in 1969 [7]. The AI was further developed into the Speech Intelligibility Index (SII) and lead to an ANSI standard for intrusive intelligibility assessment in 1997 [8]. The SII evaluates the effects of degradations in a number of frequency bands, weighted by their importance to speech intelligibility and quantifies the proportion of the speech signal that is audible to the listener. The SII score is monotonically related to the intelligibility of the signal and is given in the range 0 to 1 (where a score of 0.5 means that half of the speech cues are audible and usable to the listener) [71]. The SII describes a number of frequency band importance functions based on different speech material, which are weighting functions applied to the signal spectrum based on the importance of the particular frequency band to speech intelligibility. The SII also accounts for the effects of different types of frequency masking [8]. A number of methods have been proposed as further developments of the AI based methods such as the Speech Transmission Index (STI) [154, 155] with better modeling of the effects of reverberation and non-linear degradations on speech intelligibility.

Although the SII and STI methods have a high correlation with subjective intelligibility scores for a number of degradations, they have the deficiency in not being able to model the effects of noise suppression on intelligibility [157]. More recently, the Short-Time Objective Intelligibility Measure (STOI) for intrusive intelligibility assessment has been proposed which has been shown to have a high correlation with subjective intelligibility scores for noisy and noise-suppressed speech [157, 156, 158].

The low complexity speech intelligibility algorithm (LCIA) is a data-driven non-intrusive measure that has been shown to have a high per-condition correlation with subjective intelligibility scores of noisy and noise suppressed speech [145]. However, this method is only validated with a single speaker using condition averaging and a limited number of degradation conditions and is further evaluated in this chapter.

4.2.2 STOI

The STOI algorithm [157, 156, 158] was proposed in 2010 as an intrusive speech intelligibility assessment algorithm that can also predict the effects of time-frequency weighted speech, characteristic of the processing applied in a number of speech enhancement algorithms. The algorithm structure is outlined in Fig. 4.2.1 and begins by pre-processing the clean and degraded speech signals by resampling the signals with a frequency of 10 kHz and segmenting the signals into short time frames. A simple VAD is then applied on the clean signal to remove silent regions (regions where the frame energy is 40 dB lower than the maximum frame energy in the clean signal). This is followed by an FFT and grouping of the FFT bins into 15 one-third octave bands with the lowest and highest centre frequencies set to 150 and 4300 Hz respectively. This is followed by a short-time grouping (384 ms) of the clean and degraded signals to allow a short-time intermediate intelligibility measure to be computed. The degraded speech signal is subjected to a normalization to compensate for local energy differences between the degraded and clean signal. This is followed by clipping a signal-to-distortion ratio (SDR) to -15 dB to ensure that the effects of isolated time-frequency cells which are severely degraded do not bias the overall intelligibility calculation. The intermediate intelligibility measure is computed as the Pearson correlation coefficient between the two short-time signals and the mean of the measure over all time and frequency bins results in the final STOI score, which is monotonically related to subjective intelligibility. A logistic mapping function is proposed to linearize the relationship between STOI and subjective intelligibility scores as follows

$$f(\theta) = \frac{100}{1+\exp(a\times\theta+b)},$$

where θ is the STOI score, a and b are constants (set to -17.4906 and 9.6921 for English sentences from the IEEE database) [158].

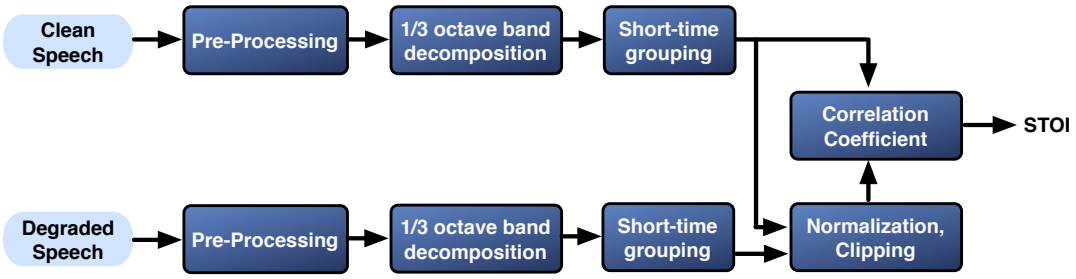

Figure 4.2.1: Overview of the STOI algorithm for intrusive intelligibility assessment [158].

4.3 Intelligibility Assessment of Noisy Speech

This section presents an investigation of non-intrusive intelligibility assessment of speech degraded by additive noise. The data-driven LCIA algorithm is proposed in Section 4.3.1, which has been published by the author in [145]. Sections 4.3.2, 4.3.3 and 4.3.4 present the LCIA2, LCIA-M and NISI algorithms respectively. The reader who is already familiar with the LCQA2, LCQA-M and NISQ algorithms (first presented in Chapter 3) may skip these sections. The databases and evaluation metrics are presented in Sections 4.3.5 and 4.3.6 respectively. This is followed by results for non-intrusive estimation of STOI scores in Section 4.3.7.

4.3.1 LCIA

The LCIA algorithm [145] is a data-driven approach for low-complexity, non-intrusive speech intelligibly assessment, developed from the LCQA algorithm [53] with a new feature, an external VAD[1], the use of a two-step feature selection and projection technique and training on databases labelled with STOI. The overall structure of LCIA is similar to the LCQA2 algorithm and begins by deriving per frame features from the speech waveform, then applying a statistical model followed by a two-step dimensionality reduction and GMM mapping. The first step is a linear prediction analysis (LPC) using 20 ms, non-overlapping frames of the speech signal. The frequency response of the LPC spectrum is used to derive a number of per frame features including the spectral flatness, spectral centroid, excitation variance and spectral dynamics. In addition, the speech variance and the iSNR per frame are computed giving a total of 6 per frame features. In addition, the first time derivatives of these (except spectral dynamics) are also computed, resulting in 11 features per frame. An external pitch estimation algorithm was not used as a feature in LCIA (although this information could be extracted from the LPC residual) due to the computational complexity of pitch tracking, and the poor correlation of this feature with subjective intelligibility scores [145]. The mean, variance, skewness and kurtosis of the per-frame features results in a 44 dimensional feature vector per utterance, which is further reduced by a correlation based feature selection and principal component analysis (PCA) based feature projection[2]. A joint GMM is trained on the projected features and the intelligibility score for each speech utterance in the training data. The GMM was tested with a number of mixtures and the optimal number of mixtures was found experimentally as the one that gave the lowest RMSE[3].

[1] as described in Chapter 2
[2] Please refer to LCQA2 description in Chapter 3 for more details.
[3] Exact numbers are given in the results section.

Feature description	Feature	Rate of change of feature
Spectral flatness	ϕ_1	ϕ_7
Spectral dynamics	ϕ_2	-
Spectral centroid	ϕ_3	ϕ_8
Excitation variance	ϕ_4	ϕ_9
Speech variance	ϕ_5	ϕ_{10}
iSNR	ϕ_6	ϕ_{11}

Table 4.1: The 11 per-frame features used in the LCIA algorithm.

4.3.2 LCIA2

This is an extension to the LCIA method presented in Section 4.3.1, with the following additional features. The additional features include pitch period (estimated using the PEFAC algorithm [52]) and the zero crossing rate. Further details can be found in Chapter 3 (Section 3.4.1), where the LCQA2 method was presented for non-intrusive speech quality assessment. The LCIA2 method differs from the LCQA2 method in the training methodology, for which a database labelled with STOI scores is used.

4.3.3 LCIA-M

This is an MFCC based development of the LCIA algorithm for non-intrusive speech intelligibly estimation, first described in Chapter 3 (Section 3.4.2) as the LCIA-M method. This method extracts 42 features per frame of the signal, whose statistics are used to obtain 168 global features. A two-step dimensionality reduction method is used to reduce the number of global features used to train a GMM. The LCIA-M method is applied here for non-intrusive speech intelligibility estimation and differs from LCIA-M in the training objective, for which a database labeled with STOI scores is used.

4.3.4 NISI

The NISI algorithm is based on the NISA framework and is aimed at speech intelligibility assessment. The 25 per-frame features are described in Table 4.2, whose mean, variance, skewness and kurtosis over all frames are used as global features.

Feature description	Feature	Rate of change of feature
Spectral flatness (LPC)	ϕ_1	ϕ_{14}
Spectral dynamics (LPC)	ϕ_2	-
Spectral centroid (LPC)	ϕ_3	ϕ_{15}
Zero crossing rate	ϕ_4	ϕ_{16}
Excitation variance	ϕ_5	ϕ_{17}
Speech variance	ϕ_6	ϕ_{18}
Pitch period	ϕ_7	ϕ_{19}
iSNR	ϕ_8	ϕ_{20}
Hilbert envelope variance	ϕ_9	ϕ_{21}
Hilbert enveloped dynamic range	ϕ_{10}	ϕ_{22}
Spectral flatness (PLD)	ϕ_{11}	ϕ_{23}
Spectral dynamics (PLD)	ϕ_{12}	ϕ_{24}
Spectral centroid (PLD)	ϕ_{13}	ϕ_{25}

Table 4.2: The 25 per-frame features used in the NISI algorithm.

Also included are 16 features characterizing the the long-term spectral deviation of the current signal from LTASS, referred to as the residual magnitude spectrum (P_{LTLD}) and mapped into 8 frequency bins, each with a bandwidth of 500 Hz. The energy in each bin as a percentage of the total energy is then computed and forms the long term features in NISI as follows

$$\phi_j = \frac{\sum_{g \in w} P_{LTLD}(g)}{\sum_{k=1}^{K} P_{LTLD}(k)},$$

where $j = [26, ..., 41]$ and w is a 500 Hz window centered on the frame of interest and the numerator is the energy of the current frame and the numerator is the total energy in the residual spectrum. The 116 global features are used to train a regression tree (CART) that maps the features to an intelligibility score. Table 4.3 presents the Spearman rank correlation coefficient for the 10 best global features[4] with STOI scores from the TNi database (described in Section 4.3.5. The iSNR feature is seen to be highly correlated with STOI (correlation of 0.89) for the TNi database and similarly, the zero-crossing rate, pitch period, LPC and PLD based features are expected to be useful in non-intrusive speech intelligibility estimation in additive noise conditions. The NISI method includes all 116 features used in the NISQ algorithm (see Section 3.4.3 for more details) but differs in the training objective, which is based on a database labeled using the STOI algorithm.

[4] correlations are calculated independently for each feature

Global feature	Correlation with STOI
$\mu(\phi_8)$	0.89
$\sigma(\phi_2)$	0.87
$\sigma(\phi_7)$	0.87
$\sigma(\phi_8)$	0.85
$\mu(\phi_2)$	0.84
$\sigma(\phi_{13})$	0.84
$\sigma(\phi_{24})$	0.83
$\sigma(\phi_{12})$	0.83
$\sigma(\phi_3)$	0.81
$\sigma(\phi_4)$	0.80

Table 4.3: Spearman rank correlations between 10 best features and STOI for the TNi database, described in Section 4.3.5.

4.3.5 Databases

This section presents the speech databases used for evaluating the non-intrusive intelligibility assessment algorithms, which are based on those used in Chapter 3 for speech quality assessment. The training and validation database is based on the additive noise partition of the TN database (referred here as the TNi database). The TNi database is thus comprised of speech from 336 speakers from the TIMIT [48] database with the addition of 15 noises from the from the NATO noise database [167] at SNR's in the range -24 to 30 dB in 3 dB steps, with speech level calculated using the P.56 method [82]. Figure 4.3.1 presents condition averaged STOI scores for each SNR in the TNi database. Additionally for the purpose of a generalization test, the additive noise partition of the C-Qual database [146] is used, comprising of car, babble and hum noise representing 21 conditions for each speaker (of which there are 4). All databases were down-sampled to 8 kHz to represent narrowband speech transmission.

4.3.5.1 Intelligibility labeling

The labelling of the databases for the purpose of speech intelligibility assessment is carried out with the intrusive STOI algorithm in a similar manner to the labelling for speech quality in Chapter 3 with the PESQ algorithm. The STOI algorithm was confirmed to give a Spearman correlation coefficient of 0.94 with subjective word intelligibility scores for the intelligibility study in [66], which is in line with the high correlations ($\geq$ 0.93) presented in [158]. Since the TNi database only contains additive noise distortions, the STOI score is expected to be well correlated with word intelligibility scores and is therefore used as the ground truth for the

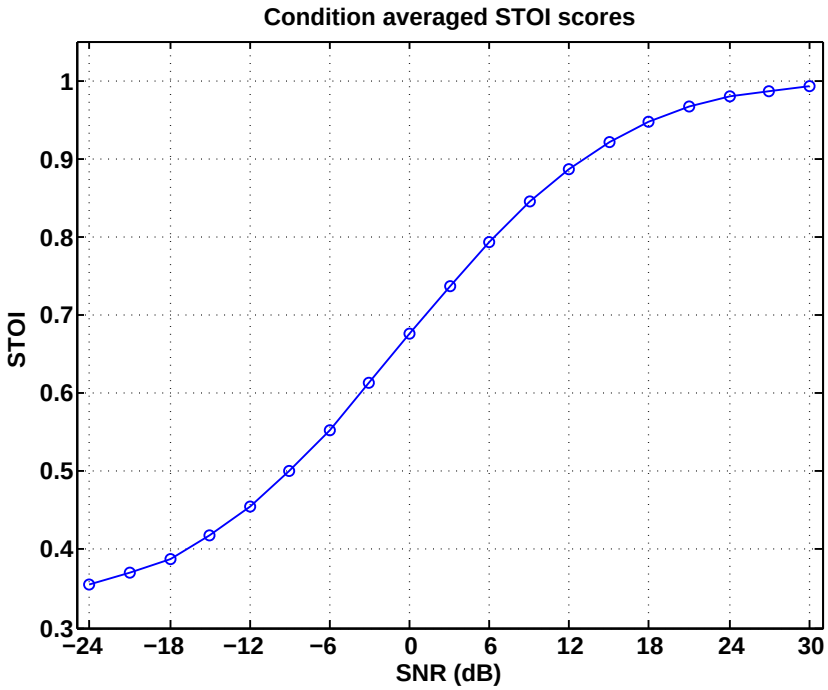

Figure 4.3.1: Condition averaged SOTI scores for each SNR in the TNi database.

non-intrusive intelligibility experiments.

4.3.5.2 Training

The training and test partitions of the TNi database contain speech from 168 speakers each with no overlap in the speech material (different text and speakers). All data-driven algorithms are trained on the TNi training partition and tested on the entire TNi test partition. The noise sources are also partitioned to ensure that the same noise samples are not used in the test as in the training.

4.3.6 Evaluation Metrics

This section defines the metrics used for measuring the performance of algorithms for speech intelligibility assessment by predicting the STOI scores in a non-intrusive scenario. The metrics outlined here are common with those used in Chapter 3 for speech quality estimation.

Spearman Correlation Coefficient (SCC)

The Spearman rank correlation coefficient is a non-parametric measure that describes the monotonic relationship between two ranked variables [104] and is calculated as the Pearson correlation coefficient between the ranked variables as follows

$$r_s(X,Y) = r_p(\check{X},\check{Y}),$$

where $\check{X}$ and $\check{Y}$ are the ranks of X and Y and r_p is the Pearson correlation coefficient. The STOI scores are monotonically related to the intelligibility of a speech signal and a logistic mapping function is applied to the STOI scores to allow direct comparison with intelligibility scores obtained in a subjective study [158]. The use of the rank correlation coefficient makes the analysis independent of the mapping function and allows a comparison of the rank similarity of the algorithms to be performed.

Root Mean Square Error (RMSE)

The root mean square error between the estimated and true scores is calculated as a measure of the estimation accuracy of each algorithm as follows

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(\varepsilon_n)^2}{N}},$$

where ε_n is the residual error defined as

$$\varepsilon_n = \theta_n - \hat{\theta}_n, \tag{4.3.1}$$

where θ_n and $\hat{\theta}_n$ are the ground truth and estimated STOI scores for the nth signal respectively.

Bin Error

The bin error evaluates the absolute mean residual error in the true and estimated STOI scores in bins of size 0.05 STOI. This measure is a simple extension of the bin error metric used for MOS and PESQ estimation, by dividing the STOI scale into 20 bins. This metric shows the percentage of signals that lie in each STOI bin and provides a view of the frequency of errors

of different magnitudes.

Two Class Classification (TCC)

This measure investigates the hit rate (HR) achieved by splitting the ground truth scores into two classes. The motivation for this metric is to evaluate the algorithms in terms of an acceptability criteria, where an acceptance threshold is set to the STOI score corresponding to 75% intelligibility, as provided by the mapping function proposed in [158] to be 0.62 on the STOI scale. The motivation for this comes from the study by K. Worrall and R. Fellows [178], where a threshold of acceptance at 75% was found to be practical. The TCC metric is also evaluated at a number of other threshold values to assess how the performance changes with different threshold values, with thresholds of 0.5 to 0.7 STOI being evaluated, corresponding to intelligiblity scores covering the range from 28% to 93%.

4.3.7 Results

This section presents the results for non-intrusive estimation of STOI scores using the databases outlined in Section 4.3.5 for the LCIA, LCIA2, LCIA-M and NISI algorithms. The results for the STOI algorithm reported here are based on the original authors' Matlab implementation. The LCIA, LCIA2, LCIA-M and NISI algorithms have been implemented in Matlab by the author. Table 4.4 shows the performance of the 4 algorithms in the estimation of STOI on the test partition of the TNi database, with the NISI algorithm outperforming the remaining algorithms, achieving a correlation (SCC) of 0.95 and an RMSE of 0.08 STOI. The NISI algorithm has a high accuracy, with 93.3% of errors less than 0.15 STOI and for an acceptance threshold of 0.62 STOI, the TCC performance is nearly 95%. The LCIA algorithm achieves a high correlation in this task (SCC = 0.91) but has a poor estimation accuracy, with an RMSE of 0.18. Similarly, the LCIA2 and LCIA-M algorithms have a low accuracy in STOI estimation. The performance of the algorithms for different acceptance thresholds is presented in Fig. 4.3.2, where the NISI algorithm can be seen to have a consistent performance with a TCC higher than 90% in the region of 0.5 to 0.7 STOI.

			Bin Error				TCC
Algorithm	**SCC**	**RMSE**	**<0.05**	**<0.10**	**<0.15**	**<0.20**	**HR**
NISI	0.95	0.08	64.0	85.4	93.3	97.0	94.7
LCIA	0.91	0.18	26.6	45.4	61.3	72.9	80.9
LCIA-M	0.88	0.15	28.3	50.9	67.1	78.9	71.1
LCIA2	0.88	0.14	18.9	40.6	64.3	83.7	91.3

Table 4.4: Non-intrusive STOI estimation performance for the TNi database.

The generalization performance of the algorithms is presented in Table 4.5, where the algorithms are trained on the TNi database[5] and tested on the additive noise partition of the C-Qual database. The performance for all algorithms in this task is much lower, with the best performance provided by the NISI algorithm (SCC of 0.86 and RMSE of 0.12). The LCIA algorithm is seen to outperform the LCIA2 and LCIA-M algorithms with and SCC of 0.82 and RMSE of 0.18.

The best performance is obtained for the LCIA2 method with 12 linear combinations of 16 features and a GMM with 9 mixtures for non-intrusive STOI estimation. Similarly, the LCIA-M performs best when 9 linear combinations of 14 best correlated features are used with a 7 mixture GMM and the LCIA algorithm performs best with 7 linear combinations of 8 features and a 7 mixture GMM. The NISI algorithm constructs a CART regression tree with 40 features,

[5] on the TNi training partition.

of which the 10 best features are presented in Table 4.6. The most important features for the task of STOI estimation of noisy speech are the mean of the iSNR, the variance of the spectral dynamics of the LPC spectrum and the variance of the spectral flatness of the PLD (long term deviation from LTASS). The best features for the LCIA method are the mean of the speech variance and iSNR based features. The iSNR is an important feature in non-intrusive STOI estimation for the TNi database and this may because this feature directly measures the amount of additive noise in the signal. Interestingly, the LCIA-M algorithm does not utilize this feature, with the important features being the variance of the velocity MFCC features.

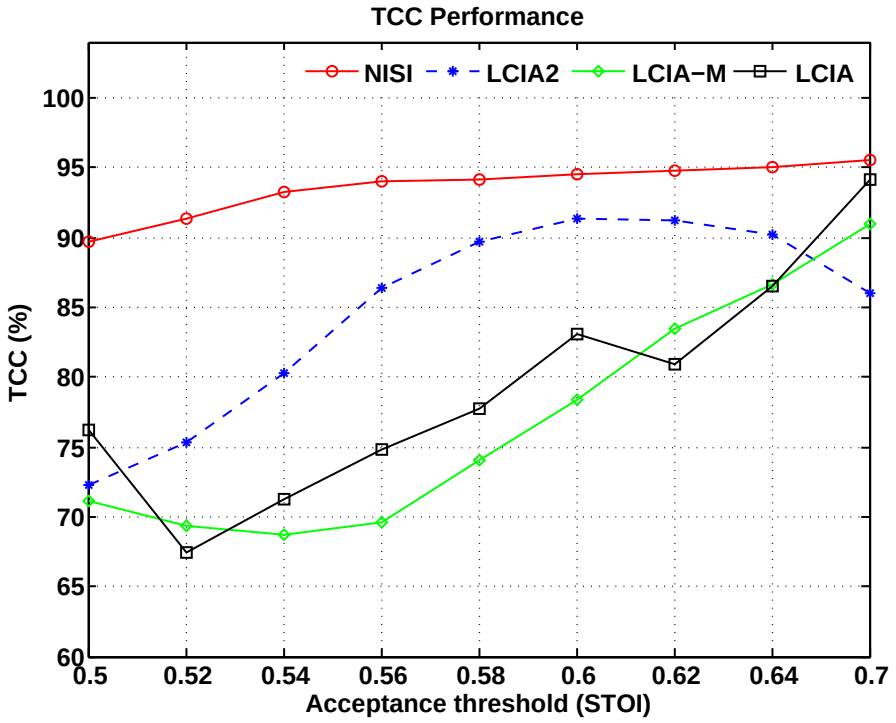

Figure 4.3.2: Results for the TCC metric using thresholds on the STOI score in the range 0.5 to 0.7 for the TNi database.

			Bin Error				TCC
Algorithm	SCC	RMSE	<0.05	<0.10	<0.15	<0.20	HR
NISI	0.86	0.12	39.3	70.2	84.5	90.5	88.1
LCIA	0.82	0.15	22.6	46.4	67.9	84.9	79.8
LCIA-M	0.48	0.42	16.7	29.8	45.2	54.8	83.3
LCIA2	0.44	0.28	10.7	16.7	32.1	48.8	72.6

Table 4.5: Non-intrusive STOI estimation performance for the additive noise partition of the CQUAL database.

Rank	LCIA	LCIA2	LCIA-M	NISI
1	$\mu(\phi_5)$	$\mu(\phi_6)$	$\sigma(\phi_{20})$	$\mu(\phi_8)$
2	$\sigma(\phi_6)$	$\kappa(\phi_{12})$	$\sigma(\phi_8)$	$\sigma(\phi_2)$
3	$\mu(\phi_6)$	$\sigma(\phi_{15})$	$\sigma(\phi_{19})$	$\sigma(\phi_{23})$
4	$\sigma(\phi_2)$	$\kappa(\phi_{13})$	$\sigma(\phi_9)$	$\mu(\phi_3)$
5	$\mu(\phi_2)$	$\gamma(\phi_2)$	$\sigma(\phi_{18})$	ϕ_{27}
6	$\sigma(\phi_1)$	$\sigma(\phi_4)$	$\sigma(\phi_1)$	ϕ_{35}
7	$\kappa(\phi_9)$	$\gamma(\phi_6)$	$\sigma(\phi_7)$	$\sigma(\phi_{13})$
8	$\kappa(\phi_6)$	$\sigma(\phi_3)$	$\sigma(\phi_{15})$	$\mu(\phi_{11})$
9	-	$\kappa(\phi_{15})$	$\sigma(\phi_{17})$	$\sigma(\phi_8)$
10	-	$\mu(\phi_2)$	$\sigma(\phi_3)$	$\kappa(\phi_{23})$

Table 4.6: The 10 best ranked features for non-intrusive STOI estimation based on the training partition of the TNi database.

4.4 Intelligibility Assessment of Noise Suppressed Speech

This section presents a preliminary investigation of non-intrusive assessment of noise suppressed speech. The STOI algorithm has been shown to have a high correlation with time-frequency processed speech [158] and is used for obtaining ground truth intelligibility scores. The remainder of this section is organized as follows, after an introduction in Section 4.4.1, the methodology and metrics are defined in Section 4.4.2 and results for non-intrusive prediction of STOI for speech that is enhanced with the spectral subtraction algorithm are presented in Section 4.4.3.

4.4.1 Introduction

One of the principal concerns of audio forensics is the enhancement of audio recordings to improve speech intelligibility [110], with US government and private laboratories conducting forensic examination of audio for speech intelligibility enhancement since the 1960s [102]. In addition, a number of speech enhancement algorithms have been developed with speech quality enhancement as the primary objective [74] and recent studies have shown that many speech enhancement algorithms deteriorate intelligibility [67, 74] even-though a positive effect on speech quality has been reported [73]. A recent investigation using three speech enhancement techniques evaluated on car and babble noise over a 12 dB SNR range showed that most algorithms deteriorate subjective word intelligibility scores [66]. Moreover, the negative effect of speech enhancement on intelligibility scores was found to be independent of the SNR [66]. The conclusion from such studies suggest that speech enhancement in the law enforcement context can have a negative impact on the intelligence value of a recording and objective methods for assessment of noise suppressed speech can play a vital role in optimizing and validating the forensic examination process.

A speech recording with poor intelligibility is likely to be rejected as admissible evidence in a court of law, however currently there is no objective criteria for validating a recording, and one solution that has been proposed in [162] is to define a percentage words correct threshold for speech recordings as an acceptability criteria in a court of law. The STOI algorithm [158] has recently been proposed for intrusive measurement of the effects of time-frequency weighted speech, which is representative of the spectral subtraction and binary mask based noise suppression methods. The spectral subtraction [16, 19] based methods rely on an estimate of the noise spectrum, which is then used to construct a time-frequency gain function which is applied to the noisy speech signal. The minimum statistics algorithm is often used to estimate the noise spectrum in speech inactive regions and track a smoothed noise power spectrum in speech

active regions after a bias compensation step [115, 113, 114]. A recent study on the effects of ideal binary mask based enhancement have reported improvements in speech intelligibility [174].

A non-intrusive intelligibility measure would allow the forensic audio examiner to validate whether the original recording met an acceptability threshold and then optimize the enhancement process and confirm that the enhanced recording was of a higher intelligibility than the original.

4.4.2 Methodology

The enhancement algorithm chosen is the spectral subtraction [19] technique used also in previous studies of the effects of processing on subjective intelligibility scores [66, 74]. The noise estimate for the algorithm in based on the minimum statistics method [115, 113, 114] with the implementation from Voicebox [21]. The TNi database is processed with the minimum statistics based spectral subtraction algorithm [16, 114, 21] and is referred to as TNi-SS database. The TNi-SS database is then labeled with the STOI algorithm, and the separation of the database into a training and test partition is maintained. Additionally, for the purpose of carrying out a generalization test, the additive noise partition of the C-Qual database is also processed by the spectral subtraction algorithm and the resulting data is labeled with STOI. The metrics from Section 4.3.6 are used for evaluating the performance of the NISI, LCIA, LCIA2 and LCIA-M algorithms.

4.4.3 Results

This section presents the results for non-intrusive assessment of STOI labelled noise suppressed speech. The results for the TNi-SS database are presented in Table 4.7 where the NISI algorithm can be seen as having the best performance with an Spearman correlation coefficient of 0.94 and an RMSE of 0.08 STOI. Moreover, 91.4% of the estimation errors are within 0.15 STOI and a TCC of 94% (at a threshold of 0.62 STOI). The performance in the TCC metric for STOI thresholds in the 0.5 to 0.7 range are presented in Fig. 4.4.1, where a consistent performance of the NISI algorithm can be seen, with a TCC of 94% +/- 2%. The same conclusion can be drawn from Table 4.8, where the NISI algorithm can again be seen to outperform the LCIA, LCIA2 and LCIA-M algorithms with an SCC of 0.89 and RMSE of 0.14 STOI. The estimation accuracy however is much lower in the generalization test for the noise-suppressed case, with 69% of the estimation errors less than 0.15 STOI.

The top 10 features for the non-intrusive assessment of noise-suppressed speech is are presented in Table 4.9, where the best feature for the NISI and LCIA algorithms is the variance of the iSNR per-frame feature. The iSNR feature is also seen to have a high importance in the assessment of the intelligibility of noise suppressed speech.

			Bin Error				TCC
Algorithm	**SCC**	**RMSE**	**<0.05**	**<0.10**	**<0.15**	**<0.20**	**HR**
NISI	0.94	0.08	61.0	82.4	91.4	95.5	94.0
LCIA	0.83	0.23	22.0	39.0	54.0	63.9	87.9
LCIA-M	0.85	0.18	22.1	43.5	60.7	72.8	69.7
LCIA2	0.89	0.19	24.4	44.6	59.9	71.9	74.1

Table 4.7: Results for non-intrusive assessment of noise-suppressed speech, labeled with the STOI algorithm for the TNi-SS database.

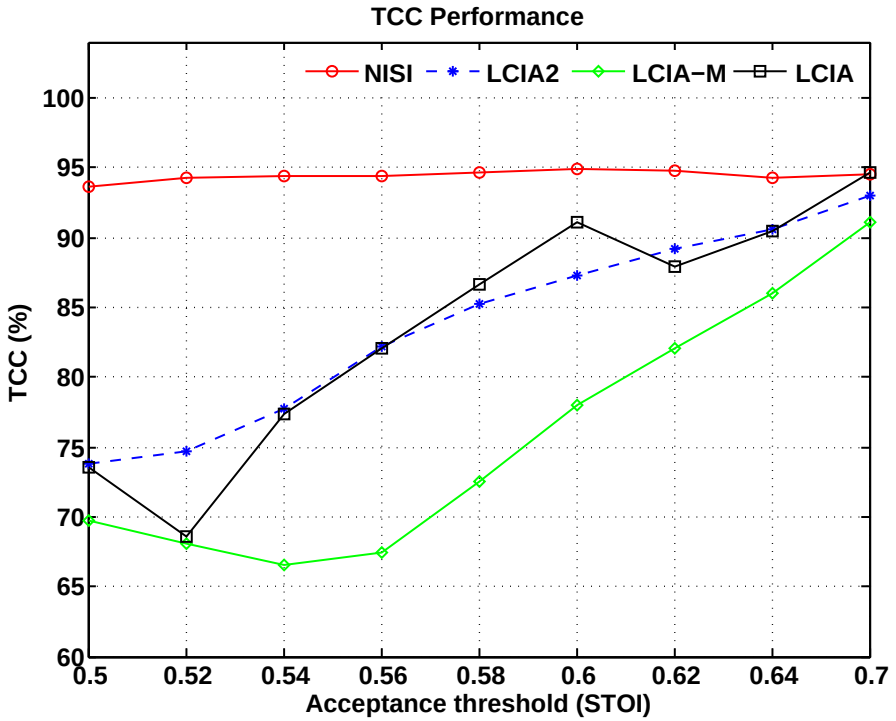

Figure 4.4.1: Performance for non-intrusive assessment of STOI for the TNi-SS database using the TCC metric with thresholds on the STOI score in the range 0.5 to 0.7.

			Bin Error				TCC
Algorithm	**SCC**	**RMSE**	**<0.05**	**<0.10**	**<0.15**	**<0.20**	**HR**
NISI	0.89	0.14	25.0	50.0	69.0	83.3	87.0
LCIA	0.83	0.17	22.6	47.6	61.9	72.6	85.7
LCIA-M	0.81	0.19	4.8	22.2	50.0	76.2	82.1
LCIA2	0.59	0.22	14.3	33.3	61.4	73.8	79.8

Table 4.8: Results for non-intrusive assessment of noise-suppressed speech, labeled with the STOI algorithm for the additive noise partition of the database.

Rank	**LCIA**	**LCIA2**	**LCIA-M**	**NISI**
1	$\sigma(\phi_6)$	$\gamma(\phi_7)$	$\sigma(\phi_{17})$	$\sigma(\phi_8)$
2	$\sigma(\phi_3)$	$\kappa(\phi_{13})$	$\sigma(\phi_8)$	$\sigma(\phi_2)$
3	$\mu(\phi_2)$	$\sigma(\phi_4)$	$\sigma(\phi_{19})$	$\mu(\phi_8)$
4	$\kappa(\phi_6)$	$\mu(\phi_2)$	$\sigma(\phi_{20})$	ϕ_{26}
5	$\gamma(\phi_6)$	$\sigma(\phi_3)$	$\sigma(\phi_1)$	$\mu(\phi_{20})$
6	$\mu(\phi_6)$	$\sigma(\phi_9)$	$\sigma(\phi_5)$	$\mu(\phi_{11})$
7	$\sigma(\phi_{11})$	$\gamma(\phi_2)$	$\sigma(\phi_7)$	ϕ_{27}
8	$\mu(\phi_{11})$	$\sigma(\phi_{15})$	$\sigma(\phi_3)$	ϕ_{39}
9	-	$\sigma(\phi_2)$	$\sigma(\phi_{18})$	$\kappa(\phi_1)$
10	-	$\mu(\phi_7)$	$\sigma(\phi_9)$	$\sigma(\phi_{25})$

Table 4.9: The 10 best ranked features for non-intrusive assessment of noise-suppressed speech from the TNi-SS database.

4.5 Summary

The problem of non-intrusive speech intelligibility assessment in the law enforcement context was presented in this chapter. A review of current techniques highlighted the requirement of robust objective measures for intelligibility estimation and the STOI algorithm was identified for modeling the effects of time-frequency weighted speech on intelligibility. The non-intrusive LCIA algorithm was further validated in this chapter along with two developments of the LCQA algorithm (LCIA2 and LCIA-M). The additive noise partition of the TN database was used as the training and validation database, which was labeled using the intrusive STOI algorithm. A novel algorithm (NISI) based on the NISA framework was shown to outperform the competing methods on all metrics, achieving a correlation of 0.95 with STOI and an RMSE of 0.08 for noisy speech. Furthermore, the performance of the NISI algorithm was shown be highly consistent over a 0.2 range of STOI using the TCC acceptance threshold metric.

A preliminary study on non-intrusive intelligibility assessment of noise suppressed speech was presented in Section 4.4 where the task was to estimate the SOTI score of noise suppressed speech. The databases were processed by the spectral subtraction algorithm and labeled with STOI. The NISI algorithm was shown to give a correlation of 0.89 and an RMSE of 0.14 STOI. The best feature for the non-intrusive intelligibility assessment was found to be the importance weighted Signal to Noise Ratio.

Chapter 5

CODEC Identification and Verification

THIS chapter presents the problem of non-intrusive CODEC identification and verification in the presence of additive noise. The NICO algorithm is presented in Section 5.3 and tested with 208 degradation conditions, including car, babble and hum noise with SNR's in the range -5 to 15 dB range and thirteen CODECs. The research presented in this chapter relates to the following publication [149].

5.1 Introduction

The use of a speech CODEC is fundamental to the efficient operation of modern communication systems by allowing a transmission system to operate at a lower bit rate while maintaining a required level of speech quality [51]. Moreover, the requirements for mobile communication devices add further constraints of delay and computational complexity to the speech coding problem [23]. Many of the speech CODECs operating with narrowband speech (corresponding to a 300 to 3400 Hz bandwidth) therefore have a low perceptual quality due to the high levels of compression that must be applied for low bit rate transmission.

The problem of CODEC identification and verification has a number of applications and the presence of a particular CODEC has been shown to have adverse effects on many speech processing systems, for example, the type of CODEC used in the transmission channel has a

dominating effect on speech quality in the absence of channel artifacts [108, 151]. A study by Besacier *et al.* [122] has shown that the presence of GSM coding significantly degrades the performance of speaker identification and verification due to the low LPC order used in the CODEC. Similarly, the presence of a CODEC can have adverse effects on automatic speech recognition (ASR) performance and a number of coding parameters have been investigated for their contribution to ASR performance [69, 164].

In the context of law enforcement audio processing, it is often required to validate the collection mechanism of the audio recordings from security devices as these can be instrumental in legal cases [56] and the identification and verification of the CODEC used in the transmission channel can help authenticate the examination process. Additionally, in the wider context of audio media forensics, the detection of particular CODECs can potentially help identify forgeries of music by detecting re-compression for audio CODECs such as WMA and MP3 [44, 109]. The identification and verification algorithm is required to be non-intrusive, since the original source speech signal is not usually available.

The remainder of this chapter is organized as follows. Section 5.2 presents a review of current algorithms, followed by a description of the NICO algorithm in Section 5.3. The databases and metrics are outlined in Section 5.4 followed by the results in Section 5.5 and a summary in Section 5.6.

5.2 Review

The 1990s witnessed an almost exponential growth in speech coding standards for a wide range of networks and applications [51]. These standards can be characterized by the bandwidth supported by the coding algorithm as narrow-band (300-3,400 Hz), wide-band (50-7,000 Hz) or super wide-band (50-14,000 Hz) [119]. The following review is concerned with narrow-band speech coding.

Two common paradigms in speech coding include waveform coding and analysis-by-synthesis coding [51]. The waveform coders are designed to reproduce the time domain waveform as accurately as possible and the G.711 [83] CODEC is used in the public switched telephone network and operates at 64 kbps [51]. This is a high quality CODEC with average MOS quality ratings between 4.0 and 4.5 [51]. The analysis by synthesis methods are based on a linear prediction model and apply perceptual distortion measures to reproduce only the important characteristics of the signal [51] with examples including the LPC based GSM Full Rate (GSM-FR) CODEC [34] and the Code Excited Linear Prediction (CELP) [141] based Adaptive Multi Rate (AMR) CODEC [35], which are deployed in digital cellular networks.

An early objective method for non-intrusive identification of the CODEC present on a telephone channel was proposed by Alley [5] in 1993. The method is based on analysis of an excitation signal processed by a channel and modeled by an adaptive filter. The method uses a multilayer perceptron classifier using features derived from the adaptive filter parameters. The results outlined in [5] are for two types of CODECs in addition to a linear channel, with identification rates between 86% and 97%. Although the results are promising, there is a requirement for an excitation signal (instead of the decoded speech signal), and only three CODECs being used with no account taken of the effects of additive noise.

More recently, an algorithm by Ludwig *et al.* [108] was proposed for the classification of low and high bit rate CODECs (2 classes) with an accuracy of 97%. The method extracts a number of features from the decoded speech signal and models them using a multidimensional Gaussian classifier. This approach is validated only with clean speech and the performance for individual CODEC identification is not presented, instead the algorithm only distinguishes between CODECs with bit rates below or above 16 kbits/s.

An algorithm for GSM-FR CODEC verification is presented in [107], where the spectral properties (a characteristic attenuation in the 2400 - 3000 Hz region of the frequency magnitude response) of the decoded signal are modeled with Gaussian distributions of the quadratic coefficient of a second order polynomial obtained from training data. The proposed method has

a hit rate of 96% but it is not clear what other CODECs were present in the evaluation besides GSM-FR or what background noise conditions were evaluated. Moreover, this method only performs CODEC verification, where a binary classification decision is taken (whether GSM-FR is present or not present).

A more recent study presents a Spectral Harmonic Decomposition (SHD) based CODEC identification method which is able to identify five types of CODEC with hit rates higher than 92% [151]. This method uses the average long term noise spectrum from the SHD of the decoded signal as a feature and a simple cross-correlation based classifier is used to assign the signal to one of the noise templates obtained from training data. The method is validated only for clean speech transmission and the sensitivity of the approach to additive noise is not shown. The algorithm proposed by Jenner *et al.* [94] extends this approach of a correlation based classifier and noise template based feature extraction. In this approach, the noise templates are generated by passing the training speech through each of the CODECs and extracting the long term average of the magnitude spectrum using FFT analysis. The time domain histogram is also used as a feature to detect the quantization present in various CODECs. The results presented are validated using clean speech material from the TIMIT database and 6 different CODECs (with 19 bit rates) with hit rates higher than 88%.

5.3 NICO

The NICO algorithm is an example application of the NISA framework for CODEC identification and verification with additional features and a CART classifier. The method begins with short-time segmentation of the speech signal into 20 ms non-overlapping frames from which an 82 dimensional per frame feature vector is extracted. This feature vector includes the features proposed in the NISA framework in Chapter 2, as well as the following features. The 10th order LPC coefficients are mapped to their Line Spectrum Frequency (LSF) representations. LSFs are a transformation of the LPC coefficients that guarantee a stable representation of the LPC model after quantization and have been successfully used in a number to speech processing applications such as speech coding [163, 33] and speech/music discrimination [95]. Additionally, 12th order Mel-Frequency Cepstral Coefficients (MFCC)s along with the velocity and acceleration features are computed using FFT motivated by a previous study on the combined use of LPC and MFCC's for speech recognition [4]. The resulting 82 per-frame features as summarized in Table 5.1. These are characterized by their mean, variance, skewness and kurtosis, resulting in 328 global features. Additionally, 16 features characterizing the long-term spectral deviation are calculated (as in Section 3.4.3), resulting in 344 global features, which are used to train a CART classification tree along with the class labels for the training data. The NICO algorithm has been implemented in Matlab by the author with the pitch estimation algorithm (PEFAC) from Voicebox [21].

Feature description	Feature	Rate of change of feature
LSF coefficients	$\phi_{1:10}$	$\phi_{24:33}$
Spectral flatness (LPC)	ϕ_{11}	ϕ_{34}
Spectral dynamics (LPC)	ϕ_{12}	ϕ_{35}
Spectral centroid (LPC)	ϕ_{13}	ϕ_{36}
Zero crossing rate	ϕ_{14}	ϕ_{37}
Excitation variance	ϕ_{15}	ϕ_{38}
Speech variance	ϕ_{16}	ϕ_{39}
Pitch period	ϕ_{17}	ϕ_{40}
iSNR	ϕ_{18}	ϕ_{41}
Hilbert envelope variance	ϕ_{19}	ϕ_{42}
Hilbert envelope dynamic range	ϕ_{20}	ϕ_{43}
Spectral flatness (PLD)	ϕ_{21}	ϕ_{44}
Spectral dynamics (PLD)	ϕ_{22}	ϕ_{45}
Spectral centroid (PLD)	ϕ_{23}	ϕ_{46}
Mel-Frequency Cepstral Coefficients	$\phi_{47:82}$	-

Table 5.1: The 82 per-frame features used in the NICO algorithm.

5.4 Methodology

This section outlines the methodology used for evaluating the NICO algorithm, beginning with a description of the database used for training and testing, followed by a description of the metrics used for performance evaluation for different types of detection criteria. As described in the review presented in Section 5.2, the current methods are either validated on a limited number of CODEC's or in more advanced studies, an ideal scenario of clean speech is assumed. This chapter proposes a non-intrusive CODEC identification and verification algorithm that is robust to additive noise.

5.4.1 Database

The database used for the experiments is based on speech from the TIMIT database [48]. The TIMIT database contains speech from American English speakers representing various accents. The speech material from TIMIT is grouped into 16 base conditions, which includes 1 condition to represent clean speech (no additive noise) and the remaining 15 conditions are obtained by adding car, babble and hum noise to the speech at SNRs of 15, 10, 5, 0 and -5 dB. The base conditions are then processed by the 13 coding systems described below,

- Linear PCM - representing uncompressed data transmission at 8 kHz sample rate and 16 bit linear quantization.
- GSM-FR [34] - representing baseline mobile transmission at a bit rate of 13 kbits/s.
- AMR [35] - representing mobile transmission at the following bit rates: 4.75, 5.15, 5.90, 6,70, 7.40, 7.95, 1020 and 12.20 kbits/s.
- G.711 A-law [83] - representing typical infrastructure routed transmission at a bit rate of 64 kbits/s.
- GSM transcoding - an example transcoding scenario where GSM to GSM communication is routed through infrastructure, which is typically using a G.711 CODEC. The signal is first coded by GSM-FR, then decoded to linear PCM, processed by G.711, decoded to linear PCM, coded to GSM-FR and then decoded to linear PCM (GSM-G711-GSM).
- Mp3 [1] - an example low bit rate Mp3 CODEC operating at 16 kbits/s, representing speech recorded by a portable speech recorder.

The database thus contains 208 degradation conditions (13 CODECS $\times$ 16 base conditions).

5.4.1.1 Training

The database is partitioned into a test and train partition, each containing 34,944 audio files representing all 208 conditions applied to speech from 168 speakers. The noise sources in the test and train partitions are separate (separate recordings) to ensure the classifier is not trained with the same noise source as that in the test set.

5.4.2 Metrics

This section outlines the three metrics used to evaluate the performance of the NICO algorithm in the tasks of CODEC verification and identification. Additionally, the concept of a confusion matrix is used to highlight more clearly where the miss-classification errors occur.

Hit Rate (HR)

The hit rate is defined as the percentage of utterances correctly classified, calculated as follows

$$\text{HR} = \frac{\sum_{n=1}^{N} \Upsilon(\tilde{\theta}_n, \theta_n)}{N} \times 100, \tag{5.4.1}$$

where $\tilde{\theta}_n$ is the estimated class label according to some detection criteria (i.e. CODEC identification) and θ_n is the actual class label for the n^{th} speech utterance. The total number of utterances in the test set is N and $\Upsilon(a,b)$ is an index function defined as:

$$\Upsilon(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}. \tag{5.4.2}$$

False Positive Rate (FPR)

The false positive rate is defined as the percentage of utterances that have been falsely classified as belonging to particular class x, calculated as follows

$$\text{FPR} = \frac{\sum_{n=1}^{N} \overrightarrow{\Upsilon}(\tilde{\theta}_n, \theta_n, x)}{N} \times 100, \tag{5.4.3}$$

where $\tilde{\theta}_n$ is the estimated class label and θ_n is the ground truth label for the n^{th} utterance and $\overrightarrow{\Upsilon}$ is an index function defined as,

$$\overrightarrow{\Upsilon}(a,b,x) = \begin{cases} 1 & \text{if } a = x \text{ and } b \neq x \\ 0 & \text{otherwise} \end{cases} . \tag{5.4.4}$$

False Negative Rate (FNR)

The false negative rate is the percentage of utterances that were not detected by the algorithm and calculated as follows

$$\text{FNR(x)} = \frac{\sum_{n=1}^{N} \overleftarrow{\Upsilon}(\tilde{\theta}_n, \theta_n, x)}{N} \times 100, \tag{5.4.5}$$

where $\tilde{\theta}_n$ is the estimated class label and θ_n is the ground truth label for the n^{th} utterance and $\overleftarrow{\Upsilon}$ is an index function defined as,

$$\overleftarrow{\Upsilon}(a,b,x) = \begin{cases} 1 & \text{if } a \neq x \text{ and } b = x \\ 0 & \text{otherwise} \end{cases} . \tag{5.4.6}$$

5.5 Results

5.5.1 CODEC Verification

This section presents the results for CODEC verification, where the task is to verify the presence of a particular CODEC. The verification is performed for each type of CODEC by training the NICO algorithm with binary labels (class present or not present). Table 5.2 shows the performance of the possible binary classifiers for the database described in Section 5.4.1. The classifier in each case is trained to identify just one CODEC ($\Theta_{x,\ x}$ =[PCM, G.711, GSM-FR, AMR, MP3, TRANS]). In this configuration the performance of the NICO algorithm is very good (high hit rates and low false positive rates) due to the binary nature of the classification task with hit rates higher than 97% in all cases. Also, the false positive rate is much lower than the false negative rate which is beneficial for audio forensic applications, where a low false positive may be desirable. The best performance is achieved for the PCM, AMR and MP3 CODEC verification tasks. The G.711 verification task is the most difficult (for the database conditions tested here) as shown by the highest false positive rate and lowest hit rate, perhaps due to the high bit rate of the CODEC.

The classification tree for this task utilizes between 3 and 14 features, with the MP3 CODEC verification model particularly sparse, using only three features to perform the verification task and the mean of the 9$^{\text{th}}$ LSF coefficient being the most important feature (Table 5.3). The most important per-frame features for the CODEC verification tasks are the 9$^{\text{th}}$ LSF coefficient, speech variance and the spectral flatness of the PLD[1].

Detection criterion	HR (%)	FPR (%)	FNR (%)
Θ_{PCM}	99.7	0.0	0.3
$\Theta_{G.711}$	97.7	0.3	2.0
Θ_{GSM-FR}	97.7	0.0	2.3
Θ_{AMR}	99.4	0.1	0.5
Θ_{MP3}	99.9	0.0	0.1
Θ_{TRANS}	98.6	0.2	1.2

Table 5.2: Classification results for the CODEC verification task. The HR, FPR and FNR are given as a percentage of the total number of files in the test set of the database. The decision criteria for each verification task is shown in the first column.

[1]Power spectrum of the long-term deviation of the signal from LTASS.

Detection criterion	NFeatures	1	2	3	4	5
Θ_{PCM}	8	$\sigma(\phi_{21})$	$\mu(\phi_9)$	$\mu(\phi_{16})$	$\kappa(\phi_{17})$	$\mu(\phi_1)$
$\Theta_{G.711}$	14	$\mu(\phi_{16})$	$\mu(\phi_{19})$	$\mu(\phi_{21})$	$\mu(\phi_{14})$	$\sigma(\phi_{20})$
Θ_{GSM-FR}	12	$\sigma(\phi_{39})$	$\mu(\phi_{19})$	$\mu(\phi_{16})$	$\sigma(\phi_8)$	$\mu(\phi_{13})$
Θ_{AMR}	6	$\mu(\phi_{16})$	$\sigma(\phi_{21})$	$\mu(\phi_{15})$	$\sigma(\phi_9)$	$\mu(\phi_{16})$
Θ_{MP3}	3	$\mu(\phi_9)$	$\mu(\phi_{16})$	$\mu(\phi_{21})$	-	-
Θ_{TRANS}	13	$\mu(\phi_{16})$	$\mu(\phi_{10})$	$\mu(\phi_9)$	$\mu(\phi_7)$	$\mu(\phi_{14})$

Table 5.3: The five best features for each CODEC verification task and the number of features used in each classification tree model.

5.5.2 CODEC Identification

The model in this configuration is tasked with identifying the class of CODEC used, that is one of (PCM, GSM, G711, MP3, Transcoding or AMR) and this gives a hit rate of 96.8%. This compares favorably with previous CODEC identification algorithms (i.e. [151]). Table 5.4 presents the results for CODEC identification for clean speech conditions, following a similar format as a confusion matrix, where the left column presents the ground truth and the other columns represent the classification result. The bold numbers present the proportion of each CODEC correctly classified and the sum of each row is equal to 1.0. The best performance is seen for the AMR CODEC with hit rates of nearly 100% for all bit rates. The GSM CODEC has the lowest hit rate of 88 %, with 10 % of the GSM data being incorrectly labeled as transcoding and 2 % as G.711. This could be due to the fact that the transcoding condition has GSM coding present in it and the G.711 CODEC has a higher perceptual quality making it difficult to differentiate between these two types of coding conditions. Table 5.5 presents the results for all degradation conditions and here the G.711 CODEC is seen to be hardest to detect when additive noise is present, achieving a hit rate of 89 % with misclassification into PCM and transcoding conditions. This may be due to the high quality of the G.711 and PCM CODECs resulting in fewer noise related characteristics for the algorithm to detect.

The performance of CODEC identification for different noises and SNRs is presented in Fig. 5.5.1, where it can be seen that the NICO algorithm is robust to additive noise and interestingly, the performance in noise is higher than that for clean speech (in terms of the hit rate). A reason for this might be that the CODEC used in this study are primarily designed to code speech (except for MP3) and the analysis-by-synthesis CODECs (AMR and GSM) produce audible artifacts when high levels of additive noise are present. Moreover, there does not seem to be a large difference in performance between the three noise types. The hit rates are greater than 95% and a standard deviation of 1% in hit rate is observed over all the test conditions.

The classification tree in this task is built using 16 features presented in Table 5.6. The most important per-frame features are the speech variance, the 8^{th} and 9^{th} LSF coefficients and the signal envelope variance. The mean of the per-frame features is the most important statistical descriptor for this task.

	PCM	GSM	G711	AMR	MP3	Transcode
PCM	**0.90**					0.10
GSM		**0.88**	0.02			0.10
G711	0.07		**0.90**			0.03
AMR 4.75 kbs				**1.00**		
AMR 5.15 kbs				**1.00**		
AMR 5.90 kbs				**1.00**		
AMR 6.70 kbs	0.02			**0.98**		
AMR 7.40 kbs				**1.00**		
AMR 7.95 kbs				**1.00**		
AMR 10.20 kbs				**1.00**		
AMR 12.20 kbs				**1.00**		
MP3 16 kps	0.05				**0.95**	
Transcode						**1.00**

Table 5.4: Classification results (proportion of files in test set) for non-intrusive CODEC identification in clean speech conditions, presented in a confusion matrix type table similar to previous studies [151, 94].

	PCM	GSM	G711	AMR	MP3	Transcode
PCM	**0.94**			0.04		0.02
GSM		**0.96**				0.04
G711	0.07		**0.89**			0.04
AMR 4.75 kbs	0.01			**0.99**		
AMR 5.15 kbs				**1.00**		
AMR 5.90 kbs				**1.00**		
AMR 6.70 kbs				**1.00**		
AMR 7.40 kbs				**1.00**		
AMR 7.95 kbs				**1.00**		
AMR 10.20 kbs				**1.00**		
AMR 12.20 kbs				**1.00**		
MP3 16 kps	0.01			0.02	**0.97**	
Transcode		0.03				**0.97**

Table 5.5: Classification results (proportion of files in test set) for non-intrusive CODEC identification in a confusion matrix type table similar to previous studies [151, 94].

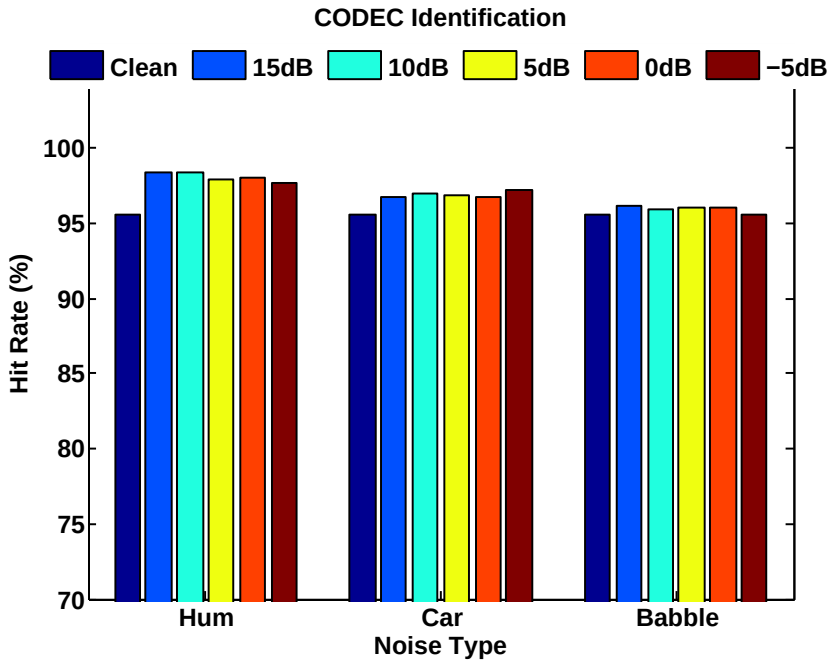

Figure 5.5.1: Results for CODEC identification with hum, babble and car noise in the -5 to 15 dB SNR range, including a clean speech condition.

Feature importance	Feature	Per-frame feature
1	$\mu(\phi_{16})$	Speech variance
2	$\mu(\phi_9)$	9^{th} LSF coefficient
3	$\mu(\phi_8)$	8^{th} LSF coefficient
4	$\mu(\phi_{19})$	Envelope variance
5	$\mu(\phi_{21})$	Spectral flatness (PLD)
6	$\mu(\phi_{15})$	LPC residual variance
7	$\mu(\phi_{14})$	Zero-crossing rate
8	$\mu(\phi_6)$	6^{th} LSF coefficient
9	$\mu(\phi_{13})$	Spectral centroid
10	$\sigma(\phi_{30})$	rate of change of 7th LSF coefficient
11	$\mu(\phi_7)$	7^{th} LSF coefficient
12	$\sigma(\phi_{57})$	11^{th} MFCC coefficient
13	$\gamma(\phi_{16})$	Speech variance
14	$\mu(\phi_2)$	2^{nd} LSF coefficient
15	$\mu(\phi_{54})$	8^{th} MFCC coefficient
16	$\sigma(\phi_2)$	2^{nd} LSF coefficient

Table 5.6: The 16 features selected for the CODEC identification task by the CART algorithm. A description of the corresponding per-frame feature is also given in the 3rd column of the table.

5.6 Summary

This chapter described a non-intrusive CODEC identification algorithm referred to as NICO that is able to identify the type of CODEC used in a communication channel with an accuracy of 96.8% for unseen data. The algorithm was tested on a database with 208 test conditions comprising three noise types (car, babble and hum) at five SNRs each, including clean speech and processed through 13 CODECs (6 different types).

The performance for CODEC verification was also presented and it was shown that NICO can perform this task with a hit rate higher than 97%. Furthermore, the false positive rate was found to be lower than the false negative rate. The speech variance and the PLD spectral flatness were found to be important per frame features for this task.

The performance for CODEC identification was presented in Section 5.5.2, where it was shown that NICO is robust to additive noise, with standard deviation in hit rate being 1% over the entire SNR range and a similar performance in hit rate for the three noise types. This is an important extension to current studies on CODEC identification, which have focused on the idealistic scenario of clean speech transmission. This is of particular importance to the law enforcement and audio forensics community, where highly degraded audio is commonplace .

Chapter 6

Speech Description Taxonomy

THIS chapter describes the Speech Description Taxonomy (SDT), which is proposed as a framework for linking various aspects of a degraded speech signal. The chapter begins with an introduction and presents a review of current literature on various aspects of the taxonomy, followed by a description of the taxonomy framework and each of its constituent components. The research presented here relates in part to the following publication [144].

6.1 Introduction

Speech signals acquired for telecommunications and surveillance applications are often degraded by the acoustic environment in which they are captured and by the non-ideal operation of the signal acquisition and transmission systems employed. The SDT is a framework for characterizing various aspects of a degraded speech signal, beginning with the degradation mechanism that models the signal processing which results in realistic degradations. The resulting degraded audio is then characterized by a vocabulary that can be used to describe the perceptual qualities of the signal. Moreover, a number of measurable and significant signal properties are extracted, allowing a data-driven framework such as NISA to extract useful metrics from the signal. The taxonomy can be implemented as a relational database allowing an audio analyst to study the interrelationships that exist between various aspects of the degraded signal. Additionally, further research in modeling of these relationships could lead to automatic diagnosis of the signal to be performed. The following outlines the SDT framework

- Degradation context: this describes the context in which a degradation occurs, for example a small car cabin might be the context for a degradation mechanism comprising of additive noise from a car engine and an Acoustic Impulse Response (AIR) of a small car cabin. This element of the SDT helps to bring together a number of different degradation mechanisms into a single 'realistic' scenario with simulated audio examples.
- Degradation mechanism: this describes the physical mechanism that produces audio with particular properties. This is the low level description of the mechanism that leads to a specific kind of degradation. Following the above mentioned example of a small car cabin context, the additive noise from a small car engine would be one degradation mechanism and the impulse response of the car cabin would be another mechanism.
- Descriptive vocabulary: this is a compact vocabulary for human description of the perceptual effects of various degradations. This would allow trained audio analysts to identify and communicate more precisely the perceptual effects of a degraded signal and may also help in performing diagnosis for speech enhancement. For the example of a small car cabin, the vocabulary class might be described as 'rumbling'.
- Signal properties: these are the measurable properties which help characterize the signal, encapsulated by features extracted from the audio signal and may be used as part of machine learning framework to perform a number of classification tasks, such as speech quality assessment, for example.

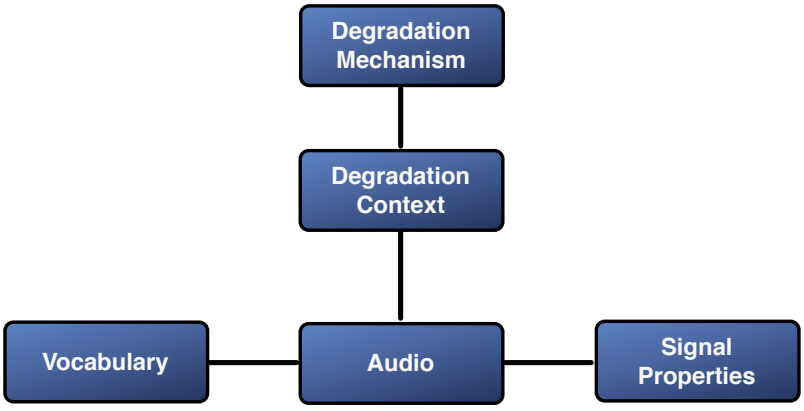

Figure 6.1.1: The speech description taxonomy framework that links various aspects of a degraded audio signal together.

The taxonomy framework is presented in Fig. 6.1.1, which shows the relationships between the various aspects of a degraded speech signal in relation to an example audio signal. The

remainder of this chapter is organized as follows, a review of various aspects of the SDT is presented in Section 6.2.

The degradation mechanism is studied in Section 6.3 and a tool for simulating complex degradation mechanisms is also presented. The development of a descriptive vocabulary is presented in Section 6.4 and the signal properties are presented in Section 6.5. This is followed by implementation details in Section 6.6 and a summary of the taxonomy in Section 6.7.

6.2 Review

6.2.1 Degradation mechanism

The ability to simulate degradation mechanisms such that the simulated speech signals have the same properties (for example the same perceptual quality and intelligibility) as a realistic recording obtained in the field of operation has a number of benefits. Such a capability allows the researcher to understand more clearly the speech acquisition and transmission system, which can provide useful insights for speech enhancement. In the law enforcement context, it is sometimes difficult to obtain large quantities of real speech recordings as they typically contain sensitive material, which implies a requirement to simulate realistic recordings for creating large databases for training and validating data-driven algorithms (such as NISQ and NISI).

The concept of a degradation mechanism simulator have been studied in the literature, for example, audio antiquating methods were presented in [165] for degrading clean audio recordings to sound antiquated. A number of processing methodologies were presented, however this study is not directed at a typical speech acquisition or transmission scenario. A recent paper [68] presents a simulation tool for simulating degradations occurring in the acoustic environment as well as those occurring over a transmission channel. However, this tool is limited in its availability and lacks the capability of incorporating an arbitrary number of degradation processes.

6.2.2 Signal properties

The detection of acoustic events based on the extraction of measurable signal properties is related to work carried in the field of audio processing for multimedia browsing and acoustic event detection. The CUIDADO project was tasked with the development of applications for audio/music content description and browsing, using signal derived features and text based descriptors [58]. A graph based music navigation algorithm is presented in [120] and a 6 class acoustic event detector using a data-driven, machine learning approach with features derived from Automatic Speech Recognition (ASR) and Music Information Retrieval (MIR) applications is presented in [12]. The area of research in audio and music retrieval has received much attention, however, at the time of writing of this thesis, the author was not aware of any studies on the application of such algorithms for law enforcement audio. The signal features derived in the SDT framework are developed from application scenarios such as speech quality and intelligibility estimation as well as CODEC identification. The classification of the

degradation types is proposed as future work in Chapter 7.

6.2.3 Speech Description Vocabulary

The creation of a concise vocabulary for describing the perceptual effects of a degraded speech signal relates closely to studies on multi-dimensional speech quality assessment. The motivation for such studies is that the quality of speech is a multi-dimensional property [116] and single dimensional quality measures do not allow for an assessment of the signal properties that lead to quality impairment [59]. An early study on multi-dimensional speech quality by McDermott [117] found speech quality to span three dimensions, interpreted as over-all clarity, a distinction between signal distortion and background distortion, and subjective loudness. The Diagnostic Acceptabiliity Measure (DAM) [169] was proposed in 1977 as a multi-dimensional speech quality framework, where the quality of the signal is characterized by 10 perceptual scales, 6 of which are for describing the perceptual properties of the foreground and 4 for describing the properties of the background. The DAM rating form presented in [130] characterizes speech quality on a scale with 16 attributes. A more recent, propriety version of DAM has been developed by Dynastat Inc which characterizes speech quality on 15 elementary perceptual qualities [143], as shown in Table 6.1.

The 1992 paper by Halka and Heute [59] presents a study on the performance of objective measures for a number of coded speech conditions, including comparison of such techniques with a 10 dimensional subjective quality experiment. The 10 attribute scales were found to be adequately represented by a two dimensional attribute-subspace using factor analysis, with the first one determined by the scales naturalness, intelligibility, clearness, nearness and rumbling. The second subspace is characterized by the scales noisiness, fullness and brightness [59].

More recently, in 2001 a study by Joseph L. Hall [60] on subjective evaluation for speech coders led to three dimensions for speech quality, labelled as naturalness, noisiness and amount of low-frequency content respectively. A comprehensive study by Mattila [116] using 170 test stimuli and more recent speech CODECs led to the development of a 21 attribute descriptive language. The 2006 study by Wältermann [172] analyses perceptual dimensions for speech quality transmitted over wide-band telephone networks, resulting in 4 dimensions identified as continuity, distance, lisping and noisiness. The multi-dimesionality of coloration perception was studies by Wen *et al.* [176] using 6 bi-polar scales labeled as warm, thin, cold, muffled, boomy and bright. It was found that two linear combinations accounted for 71% of the variance in the subjective data. The 2008 study by Wältermann [173] describes three global descriptors for speech quality as discontinuity, noisiness and coloration, with 8 additional sub-dimensions.

More recently, a paper by Sen *et al.* [143] presents an intrusive method for predicting speech quality modeling the foreground speech quality using four perceptual qualities, described as muffled, raspy, high passed and bubbly.

Diagnostic scale	Description	Aspect
SL: Low passed/Muffled	Dull, low passed, muffled	Foreground
SH: High Passed	High passed, small, distant	Foreground
ST: Thin	Thin, tinny	Foreground
SN: Nasal	Nasal, whining	Foreground
SD: Raspy	Rough, raspy	Foreground
SI: Interrupted	Interrupted, chopped	Foreground
SB: Bubbly	Babbling, slobbering	Foreground
SF: Fluttering	Fluttering, pulsating	Foreground
BNH: High Frequency Noise	Hissing, fizzing	Background
BNL: Low Frequency Noise	Rumbling, rolling	Background
BNM: Mid Frequency Noise	Rushing, roaring	Background
BB: Buzzy	Buzzing, humming	Background
BF: Bubbling	Bubbling, percolating	Background
BS: Staticy	Crackling, staticy	Background
BC: Chirping	Chirping, clicking	Background

Table 6.1: The propriety DAM elementary perceptual qualities, after [143].

6.2.4 Correspondence analysis

This section presents a review of correspondence analysis, which is a statistical tool for analyzing categorical response data [55], based on [55] [2]. Let the $M \times N$ matrix $\mathbb{X}$ represent the results from an experiment, with the explanatory variables as the columns and the response variables as the rows. Correspondence analysis is concerned with the analysis of the result matrix $\mathbb{X}$, evaluating a number of quantities for each row and column. The mass of a row reflects its importance in the sample [2], calculated as follows for row m of $\mathbb{X}$

$$r_m = \sum_n \frac{x_{m,n}}{\sum_m \sum_n x_{m,n}}. \tag{6.2.1}$$

The corresponding concept for a column is referred to as the column weight, calculated as follows for column n of $\mathbb{X}$

$$c_n = \sum_m \frac{x_{m,n}}{\sum_m \sum_n x_{m,n}}. \tag{6.2.2}$$

The vector c^{T} represents the average row profile (or centroid) of $\mathbb{X}$ and $(.)^{\mathrm{T}}$ indicates matrix transposition. The set of relative frequencies is fundamental to correspondence analysis and referred to as a profile. The profile a_{mn} of the n^{th} element of m^{th} row is defined as

$$a_{mn} = \frac{x_{m,n}}{r_m}. \tag{6.2.3}$$

In correspondence analysis, distances are measured by the chi-squared statistic (χ^2). The χ^2 distance between row m and the centroid of $\mathbb{X}$ is calculated as

$$d_{m,c} = \sqrt{\sum_{n=1}^{N} \frac{(a_{mn}-c_n)^2}{c_n}}. \tag{6.2.4}$$

The inertia of the results matrix is a measure of the variability of the row profile a_m relative to the centroid, calculated as:

$$\delta^2 = \sum_{M=1}^{M} r_m d_{m,c}^2 = \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{(a_{mn}-c_n)^2}{c_n}. \tag{6.2.5}$$

Correspondence analysis supports a hierarchical clustering method which is based on the Ward clustering algorithm [55]. The rows of the matrix $\mathbb{X}$ are successively merged, beginning with the full matrix (all rows separate) and continuing until only one row remains (all rows merged). When two rows are merged, the change in inertia of the merged matrix may be decomposed into the between-groups inertia (total inertia of the merged table) and within-groups inertia (reduction in inertia when two rows are merged). The criteria is to maximize the between-groups inertia and minimize the within-groups inertia [55]. This is equivalent to minimizing the following measure:

$$\lambda_{m,\tilde{m}} = \frac{r_m r_{\tilde{m}}}{r_m + r_{\tilde{m}}} d_{m,\tilde{m}}^2 \tag{6.2.6}$$

where r_m and $r_{\tilde{m}}$ are the row masses corresponding to the rows being merged (i.e. rows m and $\tilde{m}$) and $d_{m,\tilde{m}}^2$ is the χ^2 distance between the rows. The hierarchical clustering algorithm partitions the results matrix into a maximum of M clusters and minimum of 1.

6.3 Degradation mechanism

A key requirement for developing data-driven algorithms is large quantities of realistic speech material, that is speech signals with properties that match those of signals commonly encountered in the field of operation. It is often desirable to avoid using real speech recordings for this purpose due to security issues and difficulties in carrying out standardised and repeatable experiments. This section describes the Speech Corruption Toolkit (SCT), which is a tool for simulating realistic and repeatable degradation mechanisms. The SCT is based on a process model where the various types of corruption are simulated at the most probable point in a typical speech acquisition and transmission chain as shown in Fig.6.3.1. The SCT has been used for simulating speech with a known degradation mechanism, representing simplified operation scenarios. In practice, speech degradation mechanisms can be highly variable and such realistic scenarios are not considered in this thesis but described as future work in Chapter 7. The speech acquisition and transmission mechanisms are presented in Section 6.3.1 and a description of the software is given in Section 6.3.9.

6.3.1 Mechanism

A typical degradation chain is shown in Fig. 6.3.1, characterised by the speech acquisition, processing and transmission elements described as follows. The acoustic mixer performs the addition of a reverberant clean speech signal from the speech channel with any number of noises from the noise channel. The resultant acoustic signal is captured by a microphone, which may have its own spectral and noise characteristics. This is followed by an amplifier and signal conditioning, including band-limiting, sampling and quantisation. The digital signal may also be processed by a CODEC prior to transmission over a communication channel.

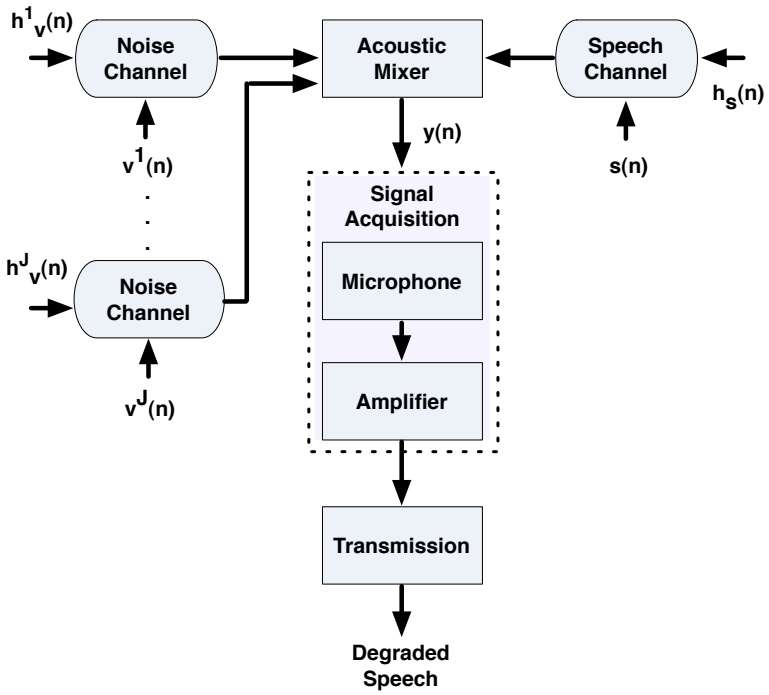

Figure 6.3.1: The degradation mechanism considered in the SCT, representing typical speech acquisition and transmission system topology.

6.3.2 Speech Channel

The speech channel is characterised by the speech-to-microphone AIR, denoted $h_s(n)$. The signal at the output of speech channel $s_a(n)$ is the convolution of the speech signal $s(n)$ with the AIR as follows

$$s_a(n) = s(n) * h_s(n). \tag{6.3.1}$$

6.3.3 Noise Channel

The noise channel simulates the convolution of additive acoustic noise with a specific noise channel impulse response (IR). Any number of noise sources $v^j(n)$ and IR $h_v^j(n)$ combinations may be present.

$$v_a^j(n) = v^j(n) * h_v^j(n). \tag{6.3.2}$$

6.3.4 Acoustic Mixer

The acoustic mixer models the addition of signals from the noise and speech channels with appropriate gain manipulation. The gain (G_v^j) for noise source j is calculated from the desired SNR in decibels and the speech level in dB Sound Pressure Level (dBSPL). In order to simulate these levels, a measurement of the active level of the source is first determined by the ITU-T P.56 standard [82] and denoted Q. The signal at the output of the acoustic mixer is defined as

$$y(n) = G_s \times s_a(n) + \sum_{j=1}^{N_j} G_v^j v_a^j(n), \tag{6.3.3}$$

where G_s is the speech gain that converts the levels to a desired sound pressure in dBSPL of a source (L) measured at 1 m, the source-microphone distance (r) metres and the reference pressure ($P_{ref} = 20\mu$ Pascals),

$$G_s = P_{ref} \times \frac{10^{(L/20)}}{\sqrt{Q} \times r}. \tag{6.3.4}$$

6.3.5 Microphone

The microphone and amplifier are part of the signal acquisition block. The intrinsic properties of the microphone may be a source of further distortion, including spectral shaping and additive noise. The microphone's impulse response can have a significant effect on the speech quality observed, as illustrated by the example response of a commercially available miniature microphone in Fig. 6.3.2. The microphone response is modelled by an FIR filter of appropriate order. If the microphone sensitivity is k mV/Pa, the voltage corresponding to the microphone's acoustic input signal is

$$V_s = k \times y(n).$$

The microphone self-noise is modeled as an equivalent noise source ν dBSPL at 1 m with corresponding voltage

$$V_{self-noise} = kp_{ref}10^{(v/20)}. \tag{6.3.5}$$

The total microphone voltage is $V_{in} = V_s + V_{self-noise}$. V_{in} is subject to the microphone's intrinsic impulse response, denoted $h_m(n)$. For simplicity, we assume that $h_m(n)$ contains its main peak at index n_d with peak duration of $2N_d$. The microphone impulse response is normalised to give

$$h'_m(n) = \frac{h_m(n)}{\sqrt{\sum_{n=-N_d/2}^{N_d/2} h_m^2(n_d+n)}}, \tag{6.3.6}$$

where N_d corresponds typically to about a 2 ms window. The microphone output voltage following convolution with $h(n)$ is

$$V_{mic}(n) = V_{in} * h'_m(n). \tag{6.3.7}$$

Let full-scale analogue deflection, dBFS, be Δ, then the microphone output, expressed in Volts and normalized in the range $[-1,1]$ is

$$V_{out} = V_{mic}/\Delta \tag{6.3.8}$$

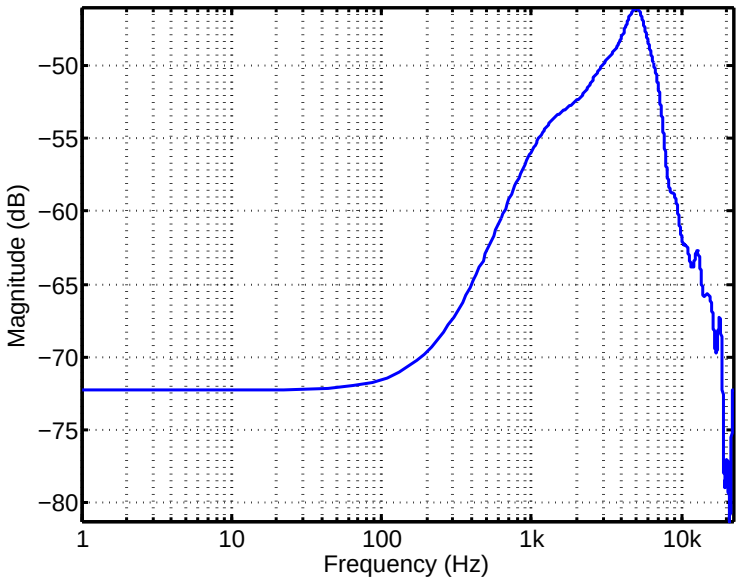

Figure 6.3.2: An example of a microphone response showing significant spectral characteristics.

6.3.6 Amplifier

Any noise or spectral shaping that is introduced by the amplifier is referred to the microphone input and included in the microphone characteristics. The first stage of the amplifier introduces a frequency-independent voltage gain. This is followed by clipping of which two types are supported: soft clipping, defined as

$$V_{soft} = \zeta \tanh(V_{in}/\zeta), \tag{6.3.9}$$

and hard clipping at amplitude ζ, defined as

$$V_{hard} = \begin{cases} V_{in} & |V_{in}| < \zeta \\ \zeta & V_{in} > \zeta \\ -\zeta & V_{in} < -\zeta \end{cases}, \tag{6.3.10}$$

where ζ is the clipping level. The SCT provides an option to set the dynamic range to be clipped for the given audio signal and scales the signal to 0 dBFS and then applies the hard clipping threshold. The signal is then rescaled to the original level. The soft clipping function tends asymptotically towards ζ, in which case hard clipping will have no effect. The signal is then resampled to the specified sampling frequency and the dynamic range set to the bit depth specified. The amplifier supports an acquisition filter specified by the two stop-band edge frequencies and the roll-off in dB/octave as shown in Fig. 6.3.3. This is implemented as an equiripple FIR filter of appropriate order.

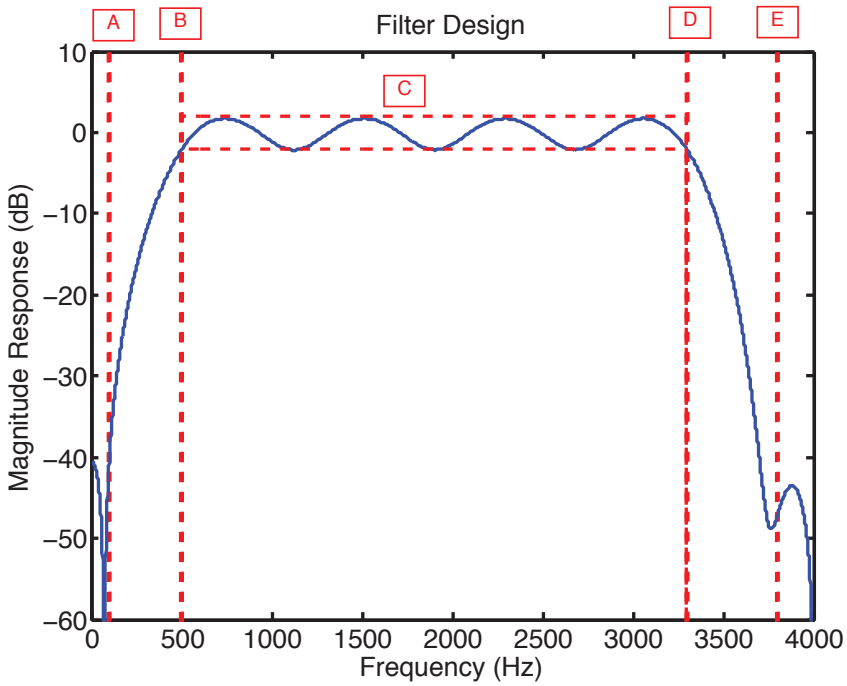

Figure 6.3.3: Acquisition filter specification. (A) is the low stop band, (B) is the low pass band, (C) is the ripple magnitude, (D) is the high pass band and (E) is the high stop band.

6.3.7 Transmission Channel

The uncompressed linear PCM signal may be transmitted over a channel with the use of CODECs, such as the Full-rate GSM [34] or AMR [35] CODEC. The SCT can model this process and supports a number of popular CODECs and also allows transcoding between any number of CODECs. Digital encoding and decoding is applied with the option of adding bit errors into the coded signal. The signal acquired at this point may be further degraded due to additional encoding for storage, which can be modelled with the batch processing mode of the SCT. The CODECs currently supported include: MP3 [1] at 16, 32, 64, 128, 160 kbps, GSM 6.10 [34], G.711 [83], GSM AMR [35] at 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20, 12.20 kbp and G.722 at 64 kbps.

6.3.8 Additional Processing

The SCT allows a number of additional processing stages to be applied including various types of filtering and non-linear distortion as described below:

- Brick-wall filtering - implemented as FIR filters to meet the specified characteristics for a low-pass, high-pass or band-pass equiripple design. This function allows bandwidth reduction to be applied at any point in the degradation chain.
- Shelf filtering - implemented as 2nd order IIR filters and are available as high or low shelf filters. The centre frequency, quality factor and gain specify the filters. This function allows for the simulation of spectral coloration that may occur due to the type of microphone used in the system.
- Clicks - temporal erasures applied at random locations in a signal. A speech activity detector is used to apply the specified number of clicks in the speech active regions with a set duration. This process also supports the simulation of drop-outs that may occur over a noisy communication channel.
- Temporal fluctuations - applied as a time varying gain function to stimulate the fluctuations in the signal level experienced in non-ideal conditions and applied at random locations in the signal and specified by the duration and number of fluctuations. This simulates the effect of a moving talker, relative to a fixed microphone resulting in a fluctuating level.

Additionally, there are a number of frequency weighting and level normalization options available in the SCT. The preparation of stimuli for subjective listening experiments often requires that the audio be normalized to a particular RMS level in dB or that a constant headroom be provided between the signal peak and digital full scale. The following options are available in the GUI and batch processing modes of the SCT

- RMS normalization method: the audio signal is normalized to a given RMS level.
- Headroom normalization method: the audio signal is normalized to provide a given headroom between the signal peak and the full scale (sets the signal peak to a specified level in dBFS).
- A-weighting: this option applies an A-weighting function [45, 21] to the signal. This function may be applied prior the RMS method for signal normalization or SNR calculation to better account for the frequency dependent effects of loudness.

6.3.9 Software

The SCT is implemented in MATLAB and the desired processing chain can be defined either via a graphical user interface (GUI) or else via a configuration text file and a command-line

interface for batch processing. The top-level GUI interface is shown in Fig. 6.3.4; clicking on the processing blocks allows their parameters to be defined. This allows the user to specify completely the degradation mechanisms in a human readable text file that allows any sequence of processes to be applied, allowing the same mechanism to be applied to an arbitrary number of audio files. Some of the supporting DSP functions are taken from Voicebox [21]. Also, some of the supporting CODEC functions rely on external libraries[1], which are freely available for research purposes.

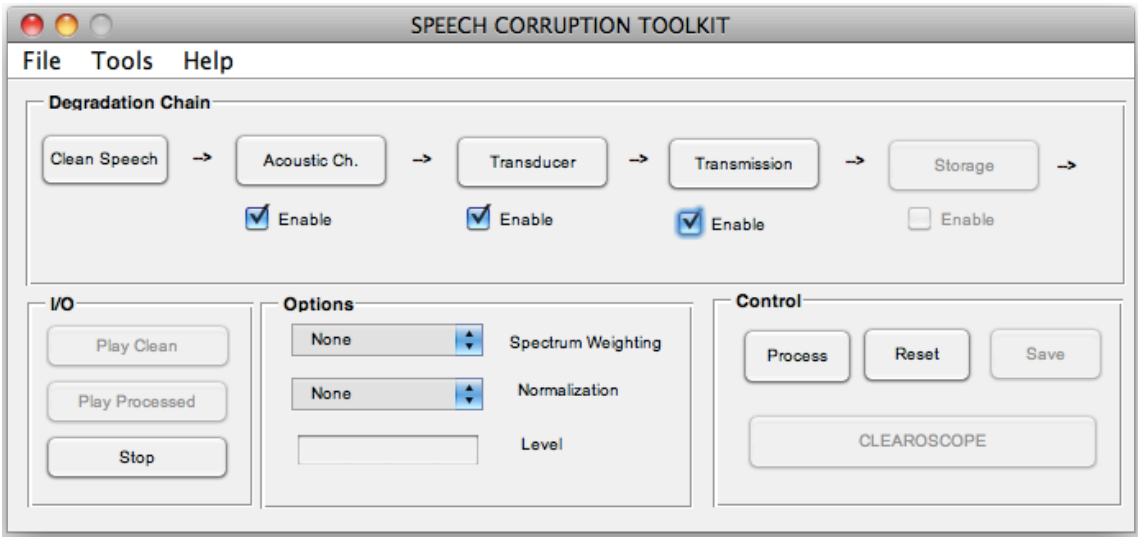

Figure 6.3.4: Speech Corruption Toolkit (SCT) graphical user interface.

[1]The GSM, G.711 and Mp3 CODECs are available as part of SoX [11] and the AMR and G.722 libraries are available as C source code that can be compiled using the mex framework in MATLAB.

6.4 Vocabulary

6.4.1 Introduction

In audio processing for surveillance and law enforcement, there is a requirement from audio analysts and practitioners for a concise vocabulary to describe the perceived characteristics of a degraded speech signal. This section presents the clustering of a large vocabulary of text descriptors into classes with similar perceptual characteristics such that words in the same class can be considered as near synonyms. This work aims at facilitating consistent and repeatable description of degraded speech such as would allow an audio transcriber to identify and communicate the audible characteristics of any degraded speech encountered. Further analysis of the signal properties of the audio associated with each class could be used to select or suggest the best enhancement regime for the corresponding degraded signals.

The scope of this study emphasizes degradations commonly found in surveillance and law enforcement audio and is limited to native English-speaking subjects. Analysis of the data using hierarchical clustering under the framework of correspondence analysis will be shown here to result in a concise vocabulary for describing the perceptual effects of the degradations and a bootstrapping validation is used to identify a robust clustering solution. As mentioned in Section 6.2.3, the current studies on this topic are concerned with integral speech quality, but here the interest is in developing a vocabulary for degraded speech description by clustering labels (rather than forming linear combinations of all labels) using a clustering approach applied to degradations that are of interest in a law enforcement scenario.

6.4.2 Methodology

This section describes the methodology used for creating the initial vocabulary, followed by a description of the database used, the subjects and the testing protocol used for the taxonomy labeling experiment (TAXIT).

Initial vocabulary and preparation for the TAXIT experiment

A first pilot study was conducted on 6 expert[2] subjects in which they were tasked to label degraded speech examples using one of 46 labels extracted from the DAM [169] and Matilla [116] studies. The subjects were asked to provide any additional labels that they would have liked to use for each of the 220 test conditions. The results from the pilot study suggested the addition of the "noisy" and "natural" labels to the entire vocabulary. The resulting 48 labels are presented using the experiment interface as shown in Fig. 6.4.2. Following this pilot study, the Taxonomy Labeling Experiment (TAXIT) was conducted using the enhanced label set and is described in the following subsections.

Database

Audio stimuli consisting of 220 sentences spoken by a male speaker were employed in the experiments [152]. A total of 55 base degraded speech conditions, denoted C01 to C55, were established, as described below:

- Brick-wall filtering (C01 - C16) : low-pass, band-pass and high-pass filters with 50 Hz transition bands and stop-band attenuation of 60 dB.
- Coloration (C17 - C30) : shelf filters with low and high cut and boost as well as 2 types of spectral tilt.
- Additive noise (C31 - C39): car, babble and hum noise were added to the speech at signal-to-noise ratios (SNR) of -5, 0 and 5 dB.
- Reverberation (C40 - C42): AIR from the MARDY database [175].
- Envelope fluctuations (C43 - C44): two random fluctuations in the speech envelope were applied.
- Clicks and dropouts (C45 - C50): temporal erasures were applied to randomly selected speech segments in the signal.
- Peak clipping (C51 - C53): symmetric hard clipping was applied with thresholds of -20, -25 and -30 dBFS.
- Clean (C54 - 55): undegraded speech.

[2]subjects with more than 4 years experience in speech and audio processing.

Each of the base conditions were processed by GSM 6.10 (13 kbps), GSM transcoding (GSM 6.10 followed by G.711 followed by GSM 6.10) and MP3 (16 kbps) CODECs and used in addition to linear PCM versions. All audio was sampled at 8 kHz and peak normalized to -10 dBFS. Further details of the database conditions are presented in Appendix C.

Subjects

A total of 51 naive, native English speaking subjects between the age of 20 and 50 years were recruited for the experiments. All subjects reported normal hearing and were paid for their participation. The listening tests were conducted in a sound-proof booth, with stimuli presented via Sennhieser HD 650 headphones driven by an RME Fireface 800 digital-to-analogue converter. Subjects received detailed instructions on the task.

Procedure

The subjects were presented with 48 text descriptors arranged on a graphical user interface according to similarity. The task was to identify the best text descriptor for "perceived quality of the audio". The presentation order of the stimuli was randomized between subjects and the average time for completing the task was 45 minutes, including a 5 minute break half way through the task. The presentation gain for the stimuli was set by the subjects at the beginning of the experiment to a "comfortable level" and all subsequent stimuli were presented to the subject at this level. The subjects were always required to select a label and were allowed to replay the audio.

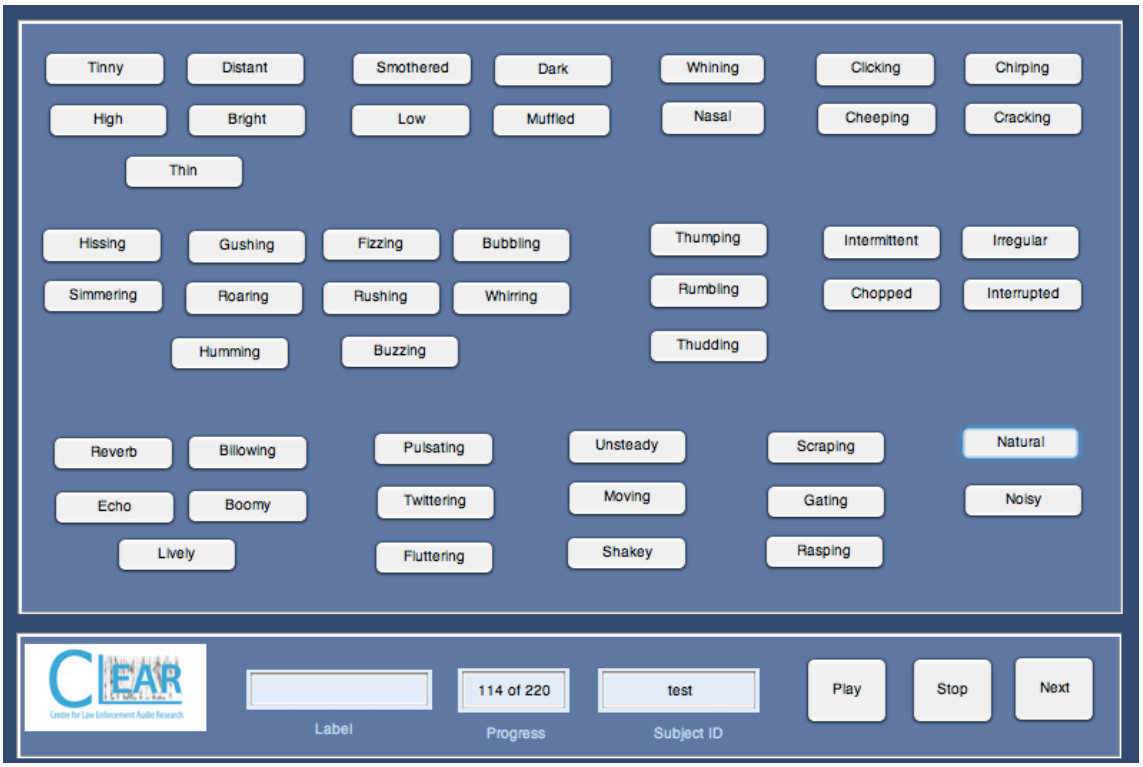

Figure 6.4.1: The TAXIT vocabulary labeling experiment's graphical user interface with the 48 labels arranged in 12 clusters.

6.4.3 Analysis of the TAXIT Experimental Results

The subjective data was analyzed using correspondence analysis. Since the objective of the experiment was to cluster the responses into classes of text descriptors representing similar perceptual characterization, a hierarchical clustering technique was applied to the data to discover the classes of vocabulary [55]. Figure 6.4.2 presents a scree plot showing the reduction in inertia achieved for different number of classes.

There are a number of methods to identify the number of classes or clusters that represent the data and a popular technique is to find the "knee" in the scree plot [24, 143]. According to the scree criterion, the results from Fig. 6.4.2 suggests that 10 clusters should be sufficient to represent the data in the TAXIT experiment. The following subsection presents a cluster validation analysis to reinforce the 10 class solution using a bootstrap technique.

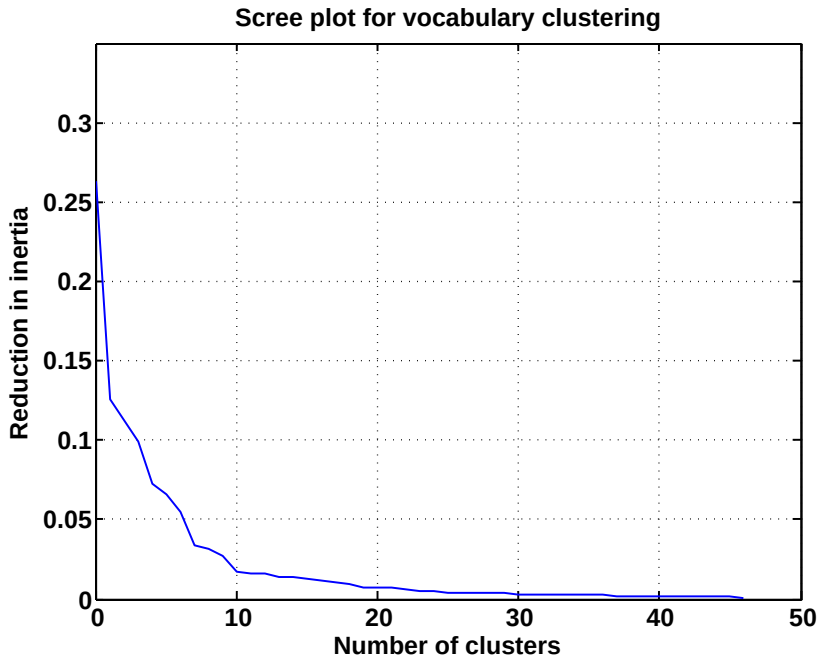

Figure 6.4.2: Scree plot for the TAXIT experiment.

Cluster Validity

Although a knee based technique is useful in identifying the number of classes, it is not guaranteed that this represents the most stable clustering solution as it may be a result of the particular sampling of the subjects from the population. A number of alternate techniques exist that exploit the data available to validate the stability of the clustering solution to variations in the data. The bootstrapping techniques construct subsamples of the data (considering the data as the population) without replacement and apply a figure of merit (FOM) to establish the reliability of the clustering solution for different number of clusters. The particular method employed is applicable to any clustering algorithm and is an example of bootstrapping cluster validity [105]. Let the number of subjects be denoted by N (N=51) and let V be the number of clusters then, τ_{ij} is an $N \times N$ connectivity matrix defined as follows:

$$\tau_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ belong to the same cluster} \\ 0 & \text{otherwise.} \end{cases} \tag{6.4.1}$$

The method proceeds by creating m ensembles of τ_{ij} using $f \times N$ subsets, where f represents a dilution factor (set to $2/3$, resulting in 34 subjects per subset[3]). The plot of the average FOM (1000 resamples) against the number of clusters is presented in Fig. 6.4.3. The local maximum of the FOM at 10 clusters suggests that a 10 cluster solution is robust to sub-sampling variations.

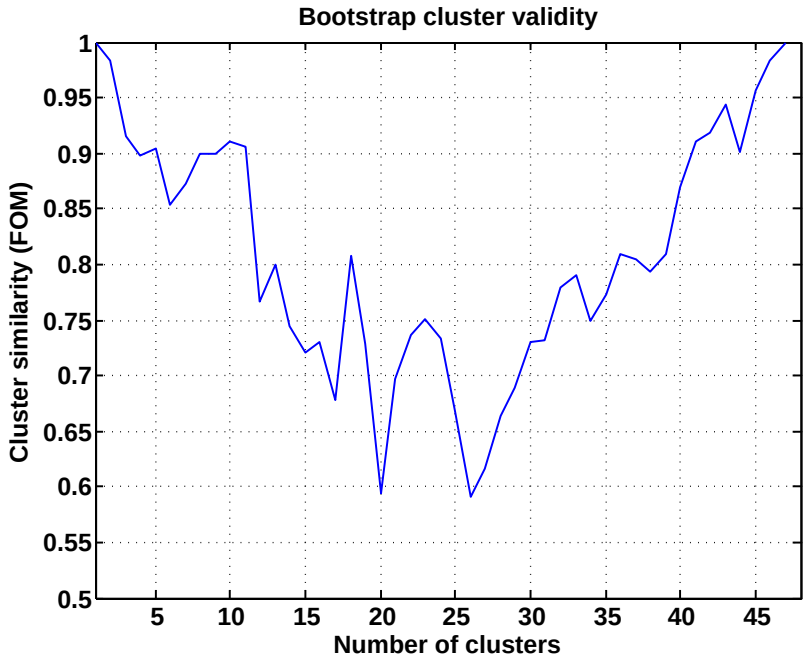

Figure 6.4.3: Average figure of merit for the hierarchical clustering algorithm applied to the TAXIT data.

Principal descriptors

The previous sections presented methods for determining and validating the number of clusters in the data. It is also desirable to determine an appropriate descriptor for each cluster based on some criterion. The first criterion chosen is the minimum χ^2 distance from each word in a cluster to its centroid, referred here as the Type A criterion. This measure identifies the principal descriptor as the label that is closest to the centroid of the cluster. Another possible criterion is the minimum of the ratio of the within class to the between class χ^2 distance for each word in a cluster. This metric is referred as the Type B criterion and identifies the principal descriptor as the label that is most distinct. The resulting principal descriptors are shown in Table 6.3.

[3] as suggested in [105]

6.4.4 Results

The analysis of the vocabulary assignment by 51 naive listeners results in the clustering of the responses into a 10 class vocabulary as shown in Fig. 6.4.2. The knee in the scree plot suggests that 10 classes in the data are sufficient and account for 95% of the inertia in the data. A cluster validation also confirms that the 10 class clustering solution is stable as shown in Fig. 6.4.3. In addition to the clustering of the vocabulary, results for two criteria for describing the principal descriptors for each class are shown in Table 6.3. The type A principal descriptors for the TAXIT degradations are shown in Table 6.4, obtained by selecting the maximum score for each of the 10 classes per condition. This labeling allows one to describe the perceptual effects of the degradations, for example, low pass filtered speech is perceived as "Muffled" when transmitted through a PCM channel, and changes to the "Pulsating" class of descriptors when a GSM CODEC is present.

Cluster	Member vocabulary
1	Cracking
2	Buzzing
3	Natural
4	Noisy
5	Humming, Boomy
6	Muffled, Smothered, Low
7	Interrupted, Intermittent, Chopped
8	Hissing, Fizzing, Moving, Rumbling, Billowing,Roaring, Gushing, Rushing
9	Tinny, Distant
10	Bubbling, Cheeping, Chirping, Irregular,Unsteady, Whirring, Echo, Reverb, Pulsating, Whining

Table 6.2: Clustered vocabulary classes.

Cluster Number	Principal Descriptor	
	Type A	Type B
1	Cracking	Cracking
2	Buzzing	Buzzing
3	Natural	Natural
4	Noisy	Noisy
5	Humming	Humming
6	Muffled	Muffled
7	Interrupted	Interrupted
8	Rushing	Hissing
9	Distant	Tinny
10	Pulsating	Cheeping

Table 6.3: Principal descriptors for the 10 classes.

Degradation	PCM	GSM	Transcoding	MP3
Low pass	Muffled	Pulsating	Rushing	Pulsating
High pass	Pulsating	Pulsating	Rushing	Pulsating
Band pass	Distant	Pulsating	Rushing	Pulsating
Low cut shelf	Natural	Pulsating	Rushing	Pulsating
Low boost shelf	Muffled	Muffled	Muffled	Pulsating
High cut shelf	Muffled	Pulsating	Rushing	Pulsating
High boost shelf	Natural	Pulsating	Pulsating	Pulsating
Spectral tilt (c/w)	Natural	Natural	Natural	Natural
Spectral tilt (ac/w)	Natural	Natural	Natural	Natural
Hum noise	Humming	Humming	Pulsating	Buzzing
Car noise	Rushing	Rushing	Rushing	Pulsating
Babble noise	Noisy	Noisy	Noisy	Noisy
Reverberation	Natural	Pulsating	Rushing	Pulsating
Envelope fluctuations	Pulsating	Pulsating	Rushing	Pulsating
Clicks	Cracking	Pulsating	Rushing	Pulsating
Dropout	Interrupted	Interrupted	Rushing	Pulsating
Peak clipping	Pulsating	Buzzing	Rushing	Pulsating
Clean	Natural	Pulsating	Rushing	Pulsating

Table 6.4: Type A principal descriptors for TAXIT database by identifying the maximum score from the 10 classes.

6.5 Signal Properties

The signal properties aspect of the Taxonomy is concerned with extracting measurable properties of the signal in a non-intrusive framework for performing speech assessment tasks, such as speech quality and intelligibility estimation and CODEC identification and verification. A number of speech assessment tasks were discussed in the preceding chapters of this thesis using the non-intrusive, data-driven NISA framework. The 82 per-frame signal features outlined in Table 6.5 cover all features used in the NISQ[4], NISI[5] and NICO[6] algorithms. The LPC, iSNR, zero crossing rate, Hilbert envelope and PLD based features were found to be particularly useful for modeling the effects of various degradations on speech quality and intelligibility (Chapters 3 and 4 respectively). The LSF and MFCC based features were included in the NICO method in Chapter 5, where they proved particularly useful for the problem of non-intrusive CODEC identification and verification.

The mean, variance, skewness and kurtosis of the 82 per-frame features results in 328 global features. In addition, 16 features for characterizing the long-term spectral deviation are calculated (as in Section 3.4.3), resulting in 344 global features. These may be used to train and model the effects of degradations on speech quality, intelligibility and CODEC identification and verification tasks. The SDT framework accommodates the addition of new features to the signal properties table as required and the particular machine learning algorithm used for modeling the features is also open to further research.

[4] see Section 3.4.3
[5] see Section 4.3.4
[6] see Section 5.3

Feature description	Feature	Rate of change of feature
LSF coefficients	$\phi_{1:10}$	$\phi_{24:33}$
Spectral flatness (LPC)	ϕ_{11}	ϕ_{34}
Spectral dynamics (LPC)	ϕ_{12}	ϕ_{35}
Spectral centroid (LPC)	ϕ_{13}	ϕ_{36}
Zero crossing rate	ϕ_{14}	ϕ_{37}
Excitation variance	ϕ_{15}	ϕ_{38}
Speech variance	ϕ_{16}	ϕ_{39}
Pitch period	ϕ_{17}	ϕ_{40}
iSNR	ϕ_{18}	ϕ_{41}
Hilbert envelope variance	ϕ_{19}	ϕ_{42}
Hilbert envelope dynamic range	ϕ_{20}	ϕ_{43}
Spectral flatness (PLD)	ϕ_{21}	ϕ_{44}
Spectral dynamics (PLD)	ϕ_{22}	ϕ_{45}
Spectral centroid (PLD)	ϕ_{23}	ϕ_{46}
MFCC	$\phi_{47:82}$	-

Table 6.5: The 82 per-frame features used for characterizing the signal properties of a degraded speech signal.

6.6 Implementation

The SDT can be implemented as a relational database [26], which allows the various aspects of the SDT to be linked together in a structured and efficient manner. The database stores the key SDT attributes as primary tables, as follows:

- Degradation context - this tabulates a number of contexts for the degradations.
- Degradation mechanism - this tabulates the individual degradation processes that lead a particular degradation contexts.
- Vocabulary - this tabulates the 10 class vocabulary from the TAXIT experiment, with associated synonyms.
- Signal properties - this tabulates the 344 features extracted for each audio signal.
- Audio - this lists the audio examples for each degradation context along with a link to the location of the audio file.

The database then uses a number of tables for linking the various attributes together. This linking is carried out in a supervised manner such that all relationships are known beforehand. The following linking tables are defined:

- Degradation context :: Degradation mechanism - this table links each degradation context to a number of mechanisms.
- Degradation context :: Audio - this table links each degradation context to example audio signals.
- Audio :: Vocabulary - this table links the audio signals to the vocabulary.
- Audio :: Signal properties - this table links the audio signals to their extracted features.

.

6.7 Summary

This chapter described the speech description taxonomy, which is a framework for linking various attributes of a degraded signal. The degradation mechanism was studied in Section 6.3 and a software tool (the SCT) for generating degraded speech data from clean speech databases was presented. The SCT enables large quantities of realistic data to be synthesised without the need for extensive recordings in the field, which may be unavailable to researchers because of security or privacy considerations. The tool includes modelling of the acoustic degradations of noise and reverberation, electronic degradations associated with the microphone and front-end, and storage/transmission degradations associated with CODECs and non-ideal channel characteristics. A key feature of the SCT is that it provides for portability and repeatability of the degradations given that the degrading processes are fully described by a human readable configuration file. The SCT was used to synthesize 220 degradation conditions for the TAXIT experiment.

The descriptive vocabulary clustering experiment, TAXIT was conducted on 51 naive native English subjects using 220 degradation conditions relevant to the surveillance and law enforcement audio processing field. Exploratory data analysis using correspondence analysis led to a clustering of a 48 label vocabulary into 10 classes. This result was further validated by performing a bootstrapping cluster analysis using a figure of merit on 1000 resamples of the data. The result showed that the 10 class clustering solution was stable to sample fluctuations. Additionally, results for two methods of determining a label from the vocabulary to serve as the principal descriptor for each cluster were presented. The concise label vocabulary provides for the identification and communication of the audible aspects of degraded speech on a 10 label vocabulary.

Finally, the signal properties were presented in Section 6.5, where 344 features were extracted from the audio to derive a number of metrics from the signal, including non-intrusive speech quality and intelligibility estimation and CODEC identification and verification using the algorithms described in the preceding chapters of this thesis. The SDT can be implemented as a relational database and can serve as a training tool for audio analysts working in the field of law enforcement and audio forensics, as outlined in Section 6.6.

Chapter 7

Conclusions

THIS chapter concludes the thesis with a summary of the key research outcomes in Section 7.1, conclusions in Section 7.2 and pointers for further work on some of those topics in Section 7.3.

7.1 Summary

The problem of non-intrusive assessment and characterization of degraded speech in the context of surveillance and law enforcement audio was investigated in this thesis. The presence of severely degraded audio is common in the field of law enforcement, with adverse effects on the intelligibility and quality of the acquired signals, reducing their value in an investigation and leading to higher transcription costs. Moreover, many speech enhancement algorithms have a negative impact on speech intelligibility and can be detrimental if inappropriately applied. These issues were outlined in the introduction to this thesis (Chapter 1). This thesis presented a non-intrusive framework for speech assessment and was applied to speech quality and intelligibility assessment as well as CODEC identification and verification. Moreover, a general framework for characterizing various aspects of a degraded signal was also proposed as the speech description taxonomy.

A novel data-driven framework was presented in Chapter 2 (referred to as the NISA framework) along with the C-Qual speech quality database. The use of intrusive algorithms for automatically labeling large quantities of speech material was also described, enabling effect-

ive development and validation of non-intrusive algorithms. A number of novel features were also described along with an evaluation of pitch estimation in additive noise and the PEFAC algorithm was identified as a noise robust pitch estimation algorithm. Chapter 3 presented an application of the NISA framework for non-intrusive speech quality assessment (referred to as the NISQ algorithm) in the commonly used per-utterance methodology as well as an initial study on the time-varying estimation of speech quality. The NISQ algorithm was validated on a large database, labelled with the PESQ algorithm, which was validated on the C-Qual database in Chapter 2. The NISQ algorithm was shown to outperform the industry standard P.563 method as well as the LCQA method and two developments of LCQA.

The NISA framework was also applied to the task of non-intrusive speech intelligibility assessment (referred to as the NISI algorithm) in Chapter 4 by labeling a large database using the intrusive STOI algorithm. This forms a first study into data-driven non-intrusive speech intelligibility assessment (as far as the author is aware at the time of writing this thesis). The non-intrusive identification and verification of the particular speech CODEC used in the transmission channel can help authenticate the audio collection mechanism as well as help optimize other speech assessment algorithms, such as those for speaker identification. The NISA framework was thus applied in Chapter 5 for CODEC identification and verification (referred to as the NICO algorithm) of speech degraded with three types of additive noise. This forms an extension to current efforts in CODEC identification algorithms, which have only been tested with clean speech material.

Finally, a framework for characterizing various aspects of a degraded signal was presented in Chapter 6 as the speech description taxonomy. This framework encapsulates the degradation mechanism and context, the measurable and significant signal properties and a concise vocabulary for describing the perceived characteristics of a degraded speech signal. The taxonomy can be implemented as a relational database and used as a tool for training audio analysts. The key research outcomes of this thesis are outlined in more detail follows.

Speech Description Taxonomy (SDT)

The SDT was presented in Chapter 6 as a framework for characterizing various aspects of a degraded speech signal and can be implemented as a relational database, allowing audio analysts to study the interrelationships that exist between various aspects of the degraded signal and further research in modeling these relationships could lead to automatic diagnosis of the signal to be performed. The SDT attributes are described in more detail as follows.

Vocabulary

This attribute of the SDT aims at providing a compact vocabulary for human description of the perceptual effects of various degradations, allowing trained audio analysts to identify and communicate more precisely the perceptual effects of a degraded signal and may additionally help in performing diagnosis for speech enhancement. This work is related to studies on multi-dimensional speech quality assessment, where the aim is to investigate the dimensions on which speech quality can be measured. The aim in the TAXIT experiment (as described in Section 6.4) was to emphasize degradations commonly found in surveillance and law enforcement audio and apply a clustering approach to identify those vocabulary items that combine in the same cluster and can be considered as near synonyms. Responses from 51 subjects were analyzed using a hierarchical clustering approach from the field of correspondence analysis. The clustering of a 48 label vocabulary resulted in 10 classes using the scree criterion for the 220 types of degradations investigated, which was verified by a bootstrap cluster validation technique using 1000 resamples of the response data. The 10 class vocabulary facilitates the identification and communication of the audible aspects of degraded speech.

Mechanism

The degradation context is an attribute that describes the context in which a degradation occurs and can combine a number of degradation mechanisms to produce a complex degradation scenario. The SCT was presented as a tool for simulating degradation mechanisms using a typical speech acquisition, processing and transmission model with the capability of applying the same degradation mechanisms to a number of 'clean' speech material. This enables large quantities of realistic data to be synthesized without the need for 'real' recordings acquired in the field of operation (access to which may be restricted due to security or privacy considerations). Moreover, this allows the researcher to investigate the mechanisms responsible for particular degradations and can provide pointers for improved audio collection and processing strategies to be adopted.

Signal properties

The signal properties attribute of the SDT is concerned with the extraction of features from the signal that can be used for speech assessment tasks. The NISA framework was presented in Chapter 2 as a data-driven framework for non-intrusive speech assessment. The NISA framework extracts a number of per-frame features and models them using the mean, variance, skewness

and kurtosis operators to form the set of global features. In addition, a number of long-term features are derived using the long-term average speech spectrum. This novel feature set is based on extracting features from the deviation in the power domain of the long term signal spectrum from LTASS. Also, novel features using the signal envelope were included along with a noise robust pitch estimation algorithm. A comparison of four pitch estimation algorithms revealed the the PEFAC algorithm had a good performance at low SNR conditions. The total feature vector comprises of 344 global features per speech signal and can model the quality and intelligibility of the signal as well as help identify the CODEC used in the transmission channel.

Non-intrusive assessment

The NISA framework was applied for a number of non-intrusive speech assessment tasks as described in the following subsections.

Speech Quality

The quality of a speech signal is a measure of the perceptual effects of degradations in a speech signal and can impact the transcription efficiency in an investigation. The non-intrusive assessment of speech quality is an important issue as typically a clean reference signal is not available and thus the algorithm must estimate the quality based only on the degraded signal. The intrusive PESQ algorithm was used to label a large database comprising of speech degraded by additive noise and transmission over real telephone channels. The additive noise partition of the C-Qual database was used as a generalization test database. The LCQA algorithm was extended with additional features such as iSNR and MFCCs and the use of a two-step dimensionality reduction scheme, resulting in the LCQA2 and LCQA-M algorithms. These were shown to outperform the baseline LCQA algorithm and the industry standard P.563 algorithm for this database. The NISQ algorithm was developed from the NISA framework and shown to outperform all other methods tested, achieving a correlation of 0.90 with PESQ and an RMSE of 0.4. The iSNR was found to be the most important feature for this task. An initial study on time-varying speech quality assessment was carried out using a block extension to the C-Qual database. The NISQ algorithm gave a consistent performance, achieving an RMSE lower than 0.5 MOS and a correlation higher than 0.90 for block sizes in the 0.5 to 8.0 second range.

Speech Intelligibility

The non-intrusive assessment of speech intelligibility is a novel application of the NISA framework. The LCQA, LCQA2 and LCQA-M methods were trained on a large database comprising of additive noise and labeled with the intrusive STOI algorithm, which is highly correlated with subjective intelligibility scores. It was shown that the NISI algorithm gave the best overall performance, achieving a correlation of 0.95 with STOI and an RMSE of 0.08. The iSNR feature was found to be important for non-intrusive speech intelligibility assessment. An initial study was performed on the estimation of the effects of speech enhancement on speech intelligibility. The speech databases were processed by the spectral subtraction algorithm and then relabeled with SOTI. The NISI algorithm achieved a correlation of 0.89 with STOI and an RMSE of 0.14 for this task.

CODEC identification and verification

The NISA framework was also further extended with the addition of LSF and MFCC features and applied to the task of CODEC identification and verification. The NICO algorithm was proposed and tested on a database comprising speech with three types of additive noise in the SNR range -5 to 15 dB and processed through 6 types of CODECs, including an example transcoding condition. The NICO algorithm was able to identify the type of CODEC used in the transmission channel with an overall hit rate of 96.8% and the algorithm was seen to be robust to additive noise with a standard deviation of 1% over the entire SNR range.

In addition to the task to CODEC identification, it is sometimes necessary to perform CODEC verification, that is to identify if a particular CODEC is present or not in the transmission channel. The NICO algorithm was also evaluated in this scenario and achieved an average hit rate of 98.8%. A classifier was constructed for each CODEC verification task utilizing between 3 and 14 features. The MP3 verification classification tree was particularly small, utilizing only 3 features to detect the presence of the particular MP3 CODEC. The important features for the verification task were the 9^{th} LSF coefficient, speech signal variance and the spectral flatness of the PLD.

7.2 Conclusions

The non-intrusive assessment of speech signals acquired in the context of law enforcement is a challenging problem due to the high levels of degradation commonly encountered. The data-driven NISA framework was proposed for this task, using CART to model a number of novel features. The NISQ method was proposed, using the NISA framework, for the non-intrusive assessment of speech quality and validated on a large database comprising of additive noise and real transmission channel degradations. The intrusive, PESQ algorithm was used to label the database, allowing large quantities of data to be automatically labeled. The NISQ method was shown to outperform the industry standard P.563 and LCQA methods in per-utterance and time-varying assessment methodologies. Such technology can help improve the speech transcription process by automatically identifying sections of good quality speech in long surveillance recordings. As far as the author is aware at the time of writing of this thesis, the LCIA algorithm is a first attempt at non-intrusive speech intelligibility assessment. The NISI algorithm was also proposed, using the NISA framework, and shown to have a high correlation and low estimation error for non-intrusive estimation of the effects of additive noise and speech enhancement on speech intelligibility. The NISI method can be beneficial for optimizing the enhancement of surveillance recordings by providing an objective feedback of the intelligibility of the speech signal. This is of particular importance as a recording that has been processed by speech enhancement must be shown to not have deteriorated the intelligibility of the signal to be admissible as evidence in a court of law.

In addition to the assessment of the quality and intelligibility of speech signals, an important requirement in the law enforcement context is the validation of the collection mechanism of audio recordings from security devices. A CODEC identification and verification method, referred to as NICO, using the NISA framework was investigated and shown to give good results in both tasks using a large database comprising of additive noise and 13 CODECS. Such technology is also of importance to the wider speech processing community, as for instance, a mobile speech recognition system may be optimized if the CODEC used in the transmission channel could be identified. Other applications include optimization for speaker verification and identification systems. Current methods in the literature have only been validated on clean speech conditions and the NICO method has been shown to be robust to the additive noise conditions investigated, which is an important requirement for any practical application.

A speech description taxonomy was also presented in this thesis, which is a useful tool for investigating the relationships that exist between different attributes of a degraded speech signal. The speech corruption toolkit was developed as a tool for synthesizing realistic degradation mechanisms that can be used to applied to large quantities of speech material in a repeatable

manner. A concise vocabulary for facilitating human description of the perceptual effects of various degradations has been developed and may additionally help in diagnosis for speech enhancement. Finally, the signal properties of the taxonomy can be used for non-intrusive speech assessment tasks such as speech quality, intelligibility or CODEC identification and verification. These tools may help better train audio analysts by providing an understanding of the relationships between different attributes of a degraded speech signal and the data-driven non-intrusive NISA framework could help optimize the analysis, validation and enhancement of law enforcement audio.

7.3 Future work

This section outlines some feasible future work that extends and further validates current research, presented according to the appropriate topic area, as follows.

SDT

The SDT presented research on speech vocabulary development with the objective of facilitating the description of perceived effects of degradations. The TAXIT protocol was tested on a number of degradations common to surveillance and law enforcement. The research carried out thus far has facilitated the following ideas for further development and validation of the TAXIT experiments.

Vocabulary based quality testing

It is possible to apply the 10 class vocabulary derived from the TAXIT experiment to other speech quality databases, such as the ITU-T P.23 [77]. This would allow an evaluation of the robustness of the vocabulary to be tested on different data as well as facilitate an analysis of the mapping of the vocabulary to MOS-LQS to be studied, in a similar manner to other studies on multi-dimensional speech quality assessment, such as [143].

Explanatory analysis of TAXIT

The responses from the TAXIT experiment were analyzed using an exploratory statistical technique, which provides the clustering solution outlined in this thesis. The statistical significance of this result is not considered so far and with additional testing it may be possible to apply explanatory data analysis and obtain significance values for the experiment. An idea maybe to conduct a study on a new set of subjects using the TAXIT protocol and then apply multiblock discriminant correspondence analysis (MUDICA) [177] to calculate the significance of the results.

NISQ

The NISQ algorithm can be further developed for predicting the effects of speech enhancement on speech quality. A recent paper proposes a non-intrusive algorithm that can model the effects of noise-suppression on speech quality [123]. This could be feasible, as the NISI algorithm was shown to give a good performance in modeling the effects of spectral subtraction on speech intelligibility, it is plausible that the NISQ algorithm can be developed for this task also.

Explanatory analysis for C-Qual

As with the TAXIT data analysis, the analysis of the C-Qual database was exploratory and it would be beneficial to apply explanatory data analysis to the C-Qual data. This would allow an evaluation of the statistically significant degradation conditions to be extracted. An possibility may be to a apply repeated hypothesis tests on pairs of conditions, with a correction methodology such as the Holm-Bonferroni correction [70]. This would identify those conditions that are significantly different at a particular significance level.

Bibliography

[1] Lame mp3 encoder. [Online]. Available: http://lame.sourceforge.net/

[2] H. Abdi and L. J. Williams, *Encyclopedia of Research Design*, N. Salkind, Ed. Thousand Oaks (CA): Sage, 2010.

[3] Y. Agrawal, E. A. Platz, and J. K. Niparko, "Prevalence of Hearing Loss and Differences by Demographic Characteristics Among US Adults," *Arch Intern Med.*, vol. 168, pp. 1522 – 1530, July 2008.

[4] K. R. Aida-Zade, C. Ardil, and S. Rustamo, "Investigation of combined use of mfcc and lpc features in speech recognition systems," *World Academy of Science, Engineering and Technology*, vol. 19, pp. 74–80, 2006.

[5] D. Alley, "Automatic identification of voice band telephony coding schemes using neural networks," *Electronics Letters*, vol. 29, no. 13, pp. 1156–1157, June 1993.

[6] *Auditory Non-Intrusive Quality Estimation Plus (Anique+): Perceptual Model for Non-Intrusive Estimation of Narrowband Speech Quality*, American National Standards Institute Std. ATIS-PP-01 000 005, 2006.

[7] ANSI, "Methods for the calculation of the articulation index," American National Standards Institute, New York, ANSI Standard ANSI S3.5–1969, 1969.

[8] ——, "Methods for the calculation of the speech intelligibility index," American National Standards Institute, ANSI Standard S3.5–1997 (R2007), 1997.

[9] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 3, pp. 201–212, Jun. 1976.

[10] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2B, pp. 637–655, Aug. 1971.

[11] C. Bagwell. (2008) SoX: Sound eXchange. [Online]. Available: sox.sourceforge.net

[12] V. Barbosa, T. Pellegrini, M. Bugalho, and I. Trancoso, "Browsing videos by automatically detected audio events," in *EUROCON - International Conference on Computer as a Tool (EUROCON), 2011 IEEE*, April 2011.

[13] J. Beerends and J. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *Journal Audio Eng. Soc.*, vol. 42(3), pp. 115–123, 1994.

[14] D. Benitez, P. Gaydecki, A. Zaidi, and A. Fitzpatrick, "The use of the hilbert transform in ecg signal analysis," *Computers in Biology and Medicine*, vol. 31, no. 5, pp. 399–406, 2001.

[15] F. Beritelli, S. Casale, R. Grasso, and A. Spadaccini, "Performance evaluation of SNR estimation methods in forensic speaker recognition," in *Emerging Security Information Systems and Technologies (SECURWARE), 2010 Fourth International Conference on*, Jul. 2010, pp. 88 –92.

[16] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1979, pp. 208–211.

[17] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[19] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[20] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. CRC Press, 1984.

[21] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997.

[22] K. L. Brown and E. B. George, "CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 105–108.

[23] M. Budagavi and J. Gibson, "Speech coding in mobile radio communications," *Proceedings of the IEEE*, vol. 86, no. 7, pp. 1402–1412, jul 1998.

[24] R. B. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966.

[25] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *Proc. European Conf. on Speech Communication and Technology*, Sep. 1995, pp. 867–870.

[26] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, vol. 13, no. 6, pp. 377–387, June 1970.

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

[28] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[29] G. De'ath and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, pp. 3178 – 3192, 2000.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[31] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001.

[32] J. Egan, "Articulation testing methods," *Laryngoscope*, vol. 58(9), pp. 955–991, 1948.

[33] J. Erkelens and P. Broersen, "On statistical properties of the line spectrum pairs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1995.

[34] E. T. S. I. (ETSI), *GSM 06.10: Full Rate (FR) Speech Transcoding*, European Telecommunications Standards Institute (ETSI) Recommendation GSM 6.10, 1995.

[35] *GSM 06.90: Adaptive Multi-Rate (AMR) Speech Transcoding*, European Telecommunications Standards Institute (ETSI) Recommendation GSM 06.90, 1998.

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

[37] ——, "Feature mining for GMM-based speech quality measurement," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, vol. 2, Nov. 2004, pp. 2290–2294.

[38] ——, "Enhanced non-intrusive speech quality measurement using degradation models," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 2006, pp. I–I.

[39] ——, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Process. Lett.*, vol. 13, no. 2, pp. 108–111, Feb. 2006.

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

[41] T. H. Falk and W.-Y. Chan, "Hybrid Signal-and-Link-Parametric Speech Quality Measurement for VoIP Communications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1579 – 1589, November 2008.

[42] T. H. Falk, H. Yuan, and W.-Y. Chan, "Single-Ended Quality Measurement of Noise Suppressed Speech Based on Kullback-Leibler Distances," *Journal of Multimedia*, vol. 2, no. 5, pp. 17 – 26, September 2007.

[43] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[44] X. Feng and G. Doërr, "Fld-based detection of re-compressed speech signals," in *Proc. 12th ACM workshop on Multimedia and security*, New York, NY, USA, 2010, pp. 43–48.

[45] H. Fletcher and W. Munson, "Loudness, its definition, measurement and calculation," *J. Acoust. Soc. Am.*, vol. 5, pp. 82–108, 1933.

[46] H. Fletcher and J. Steinberg, "Articulation testing methods," *Bell Syst. Tech. J.*, vol. 8, pp. 806–854, 1929.

[47] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[48] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[49] N. D. Gaubitch, M. Brookes, and P. A. Naylor, "Blind channel identification in speech using the long-term average speech spectrum," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, Aug. 2009.

[50] N. D. Gaubitch, M. Brookes, P. A. Naylor, and D. Sharma, "Bayesian adaptive method for estimating speech intelligibility in noise," in *Proc AES Conf on Audio Forensics*, Hillerød, Denmark, Jun. 2010.

[51] J. Gibson, "Speech coding methods, standards, and applications," *Circuits and Systems Magazine, IEEE*, vol. 5, no. 4, pp. 30–49, 2005.

[52] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Aug. 2011.

[53] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

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

[55] M. Greenacre, *Correspondence Analyis in Practice*, 2nd ed. Chapman & Hall/CRC, 2007.

[56] S. Gupta, S. Cho, and C.-C. Kuo, "Current developments and future trends in audio authentication," *Multimedia, IEEE*, vol. 19, no. 1, pp. 50–59, jan. 2012.

[57] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, March 2003.

[58] P. H. H. Vinet and F. Pachet, "The cuidado project," in *Proceedings of the 3rd International Symposium on Music Information Retrieval*, France, 2002.

[59] U. Halka and U. Heute, "A new approach to objective quality-measures based on attribute-matching," *Speech Communication*, vol. 11, no. 1, pp. 15–30, 1992.

[60] J. L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *J. Acoust. Soc. Am.*, vol. 110, no. 4, pp. 2167–2182, 2001.

[61] M. Hansen and B. Kollmeier, "Continuous assessment of time-varying speech quality," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2888–2899, 1999.

[62] P. S. Hansen. Online. [Online]. Available: http://www.itu.dk/courses/TKG/E2005/Mdir/

[63] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[64] W. Hess and H. Indefrey, "Accurate pitch determination of speech signals by means of a laryngograph," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 9, 1984, pp. 73–76.

[65] U. Heute, S. Moller, A. Raake, A. Scholz, and M. Waltermann, "Integral and Diagnostic Speech-Quality Measurement: State of the Art, Problems, and New Approaches," in *Proc. Forum Acusticum*, Budapest, Hungary, 2005.

[66] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale, "Effects of noise suppression on intelligibility: dependency on signal-to-noise ratios," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 531–539, 2012.

[67] G. Hilkhuysen and M. Huckvale, "Signal properties reducing intelligibility of speech after noise," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010.

[68] H. G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. European Conf. on Speech Communication and Technology*, 2005.

[69] S. Ho Choi, "A study on speech coders for automatic speech recognition in adverse communication environments," in *Informatics Engineering and Information Science*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2011, vol. 252, pp. 67–75.

[70] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, pp. 65 – 70, 1979.

[71] B. W. Y. Hornsby, "The Speech Intelligibility Index: What is it and what's it good for?" *Hearing Journal*, vol. 57, no. 10, pp. 10–17, October 2004.

[72] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

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

[74] ——, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, pp. 1777–1786, 2007.

[75] *Methods for subjective determination of transmission quality*, Online, International Telecommunications Union (ITU-T) Recommendation P.800, Aug. 1996. [Online]. Available: http://www.itu.int/rec/T-REC-P.800/en

[76] *Modulated Noise Reference Unit (MNRU)*, International Telecommunications Union (ITU-T) Recommendation P.810, 1996.

[77] *ITU-T coded-speech database*, International Telecommunications Union (ITU-T) Supplement P.Sup23, Feb. 1998.

[78] *Subjective performance evaluation of telephone band and wideband codecs*, International Telecommunications Union (ITU-T) Recommendation P.830, 1998.

[79] *Continuous evaluation of time varying speech quality*, Online, International Telecommunications Union (ITU-T) Recommendation P.880, May 2004. [Online]. Available: http://www.itu.int/rec/T-REC-P.880/en

[80] *Single-ended method for objective speech quality assessment in narrow-band telphony applications*, International Telecommunications Union (ITU-T) Recommendation P.563, 2004.

[81] *The E-model: a computational model for use in transmission planning*, International Telecommunications Union (ITU-T) Recommendation G.107, December 2011.

[82] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[83] ——, *Pulse Code Modulation (PCM) of Voice Frequencies*, International Telecommunications Union (ITU-T) Rec. G.711, Nov. 1998.

[84] ——, *Artificial Voices*, International Telecommunications Union (ITU-T) Recommendation P.50, Sep. 1999.

[85] ——, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[86] ——, *Mapping function for transforming P.862 raw result scores to MOS-LQ*, International Telecommunications Union (ITU-T) Recommendation P.862.1, 2003.

[87] ——, *Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2*, International Telecommunications Union (ITU-T) Recommendation P.862.3, Nov. 2007.

[88] ——, *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862.2, November 2007.

[89] ——, *Perceptual objective listening quality assessments*, International Telecommunications Union (ITU-T) Recommendation P.863, 2011.

[90] M. M. J. Chen, K. K. Paliwal and S. Nakamura, "Robust MFCCs derived from differentiated power spectrum," in *Proc. EUROSPEECH*, Scandinavia, 2001.

[91] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, February 1997.

[92] A. K. Jain, R. P. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4 – 37, January 2000.

[93] C. R. Jankowski, Jr., , A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1990, pp. 109–112.

[94] F. Jenner and A. Kwasinski, "Highly accurate non-intrusive speech forensics for CODEC identification from observed decoded signals," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012.

[95] G. P. K. El-Maleh, M. Klein and P. Kabal, "Speech/music discrimination for multimedia application," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2000.

[96] D. Kalikow, K. Stevens, and L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1337–1351, 1977.

[97] L. J. Kao and C. C. Chiu, "Mining the customer credit by using the neural network model with classification and regression tree approach," in *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th*, vol. 2, 2001.

[98] D.-S. Kim, "ANIQUE: An Auditory Model for Single-Ended Speech Quality Estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 821 – 831, September 2005.

[99] W. B. Kleijn, P. Kroon, L. Cellario, and D. Sereno, "A 5.85 kbps CELP algorithm for cellular applications," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Minneapolis, USA, 1993.

[100] S. A. Klein, "Measuring, estimating, and understanding the psychometric function: A commentary," *Perception & Psychophysics*, vol. 63, no. 8, pp. 1421–1455, 2001.

[101] H. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Michigan, USA, May 1995.

[102] B. E. Koenig, D. S. Lacey, and S. A. Killion, "Forensic enhancement of digital audio recordings," *Journal Audio Eng. Soc.*, vol. 55, no. 5, pp. 352–371, May 2007.

[103] K. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, 1962.

[104] E. L. Lehmann and H. J. M. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[105] E. Levine and E. Domany, "Resampling method for unsupervised estimation of cluster validity," *Neural Computation*, vol. 13, pp. 2573–2593, 2001.

[106] W.-Y. Loh, "Classification and regression trees," *WIREs Data Mining Knowl Discov*, vol. 1, no. 1, pp. 14 – 23, 2011.

[107] T. Lugwig, "Comfort noise detection and GSM-FR-CODEC detection for speech-quality evaluation in telephone networks," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, 2002, pp. 309–312.

[108] T. Lugwig and U. Heute, "Detection of digital transmission systems for voice quality measurements," in *Proc. European Conf. on Speech Communication and Technology*, 2001, pp. 1699–1702.

[109] D. luo, W. Luo, R. Yang, and J. Huang, "Compression history identification for digital audio signal," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012.

[110] R. C. Maher, "Audio forensic examination," *IEEE Signal Processing Magazine*, vol. 26, pp. 84–94, 2009.

[111] L. Malfait, J. Berger, and M. Kastner, "P.563 - the ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, 2006.

[112] P. Manchester, "Found sound: an introduction to forensic audio," *Sound on Sound*, vol. 750, pp. 90–95, 2010.

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

[114] ——, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, Jun. 2006.

[115] ——, "Spectral subtraction based on minimum statistics," in *Proc. European Signal Processing Conf*, 1994, pp. 1182–1185.

[116] V.-V. Mattila, "Semantic analysis of speech quality in mobile communications: descriptive language development and mapping to acceptability," *Food Quality and Preference*, vol. 14, pp. 441–453, November 2003.

[117] B. McDermott, "Multidimensional analysis of circuit quality judgements," *J. Acoust. Soc. Am.*, vol. 45, no. 3, pp. 774–781, 1969.

[118] G. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, vol. 27, no. 2, pp. 338–352, 1955.

[119] S. Moller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, pp. 18–28, 2011.

[120] C. Muelder, T. Provan, and K.-L. Ma, "Content based graph visualization of audio data for music library navigation," in *Multimedia (ISM), 2010 IEEE International Symposium on*, dec. 2010.

[121] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, pp. 52–55, Jan. 2006.

[122] L. B. nad S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini, "GSM coding and speaker recognition," in *Proc. ICASSP*, vol. 2, 2000, pp. 1085–1088.

[123] M. Narwaria, W. Lin, I. McLoughlin, S. Emmanuel, and L.-T. Chia, "Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1217 – 1232, may 2012.

[124] P. A. Naylor, "Lectures on speech processing, Imperial College London." [Online]. Available: http://www.ee.ic.ac.uk/naylor/SpeechProcessing.html

[125] P. A. Naylor, N. D. Gaubitch, D. Sharma, G. Hilkhuysen, M. Huckvale, and M. Brookes, "Intelligibility estimation in law enforcement speech processing," in *Proc ITG Conf on Speech Communication*, Bochum, Germany, Oct. 2010.

[126] M. Nilsson, S. Soli, and J. Sullivan, "Development of hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[127] R. Plomp and A. Mimpen, "Speech-reception threshold for sentences as a function of age and noise level," *J. Acoust. Soc. Am.*, vol. 66, no. 5, pp. 1333–1342, 1979.

[128] P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. IEEE Intl. Conf. Pattern Recognition,*, 1994.

[129] P. Pudil, J. Novovicova, and J. Kittler, "Floating Search Methods in Feature Selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119 – 1125, 1994.

[130] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, Jan. 1988.

[131] T. Quatieri and R. Dunn, "Speech enhancement based on auditory spectral change," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)ß*, vol. 1, Florida, USA, May 2002.

[132] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice Hall, 1988.

[133] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey, USA: Prentice-Hall, 1978.

[134] L. R. Rabiner, M. J. Cheng, A. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 24, pp. 399–418, 1976.

[135] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 1996, pp. 113–116.

[136] A. Rix, "Perceptual speech quality assessment - a review," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2004, pp. 1056–1059.

[137] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[138] A. Rix and M. Hollier, "The perceptual analysis measurement for robust end-to-end speech quality assessment," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1515–1518.

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

[140] A. Rix, R. Reynolds, and M. Hollier, "Perceptual measurement of end-to-end speech quality over audio and packet-based networks," in *Audio Engineering Society Convention 106*, May 1999.

[141] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 10, 1985, pp. 937–940.

[142] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.

[143] D. Sen and W. Lu, "Objective evaluation of speech signal quality by the prediction of multiple foreground diagnostic acceptability measure attributes," *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 4087–4103, 2012.

[144] D. Sharama, G. Hilkhuysen, P. A. Naylor, N. D. Gaubitch, M. Huckvale, and M. Brookes, "Descriptive vocabulary development for degraded speech," in *Proc. Interspeech Conf.*, 2012.

[145] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Denmark, Aug. 2010.

[146] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, M. Brookes, and P. A. Naylor, "C-Qual - a validation of PESQ using degradations encountered in forensic and law enforcement audio," in *Proc. AES Conf. on Audio Forensics*, Hillerød, Denmark, Jun. 2010.

[147] D. Sharma and P. A. Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasglow, Aug. 2009.

[148] D. Sharma, P. A. Naylor, N. Gaubitch, and M. Brookes, "Short-time objective assessment of speech quality," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Aug. 2011.

[149] D. Sharma, P. A. Naylor, N. D. Gaubitch, and M. Brookes, "Non intrusive CODEC detection algorithm," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Mar. 2012.

[150] J. Shlens. (2006, November) A Tutorial on Principal Component Analysis. Online. [Online]. Available: http://www.cs.cmu.edu/~elaw/papers/pca.pdf

[151] K. Sholz, L. Leutelt, and U. Heute, "Speech-codec detection by spectral harmonic-plus-noise decomposition," in *Proc. Asilomar Conference on Signals, Systems and Computers*, 2004, pp. 2295–2299.

[152] M. W. Smith and A. Faulkner, "Perceptual adaptation by normally hearing listeners to a simulated "hole" in hearing," *J. Acoust. Soc. Am.*, vol. 120, pp. 4019–4030, 2006.

[153] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Letters to Nature*, vol. 416, pp. 87 – 90, 2002.

[154] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[155] ——, "Mutual dependence of the octave-band weights in predicting speech intelligibility," *Speech Communication*, vol. 28, pp. 109–123, 1999.

[156] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "On predicting the difference in intelligibility before and after single-channel noise reduction," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2010.

[157] ——, "A short-time objecitve intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[158] ——, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, September 2011.

[159] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier, 1995, pp. 495–518.

[160] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm for estimation of reference-quality glottal closure instants from electroglottograph signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[161] ——, "The SIGMA algorithm: A glottal activity detector for electroglottographic signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1557–1566, Nov. 2009.

[162] J. Tibbitts and Y. Lu, "Forensic Applications of Signal Processing," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 104–111, March 2009.

[163] J.-Y. Tourneret, "Statistical properties of line spectrum pairs," *Signal Processing*, vol. 65, no. 2, pp. 239–255, 1998.

[164] D. L. P. H. B. Turunen, J.; Vlaj, "Speech coding parameters and their influence on speech recognition," *Lipping, T. (Ed.) Signal processing ressearch series: comparative analysis of linear and nonlinear parametric models for speech coding, Series A - Tampere University of Technology*, vol. 42, pp. 6–10, 2003.

[165] V. Valimaki, S. Gonzales, O. Kimmelma, and J. Parviainen, "Digital Audio Antiquing Signal Processing Methods for Imitating the Sound Quality of historical Recordings," *Journal of the Audio Engineering Society*, vol. 56, no. 3, pp. 115–139, March 2008.

[166] R. J. M. van Hoesel and R. S. Tyler, "Speech perception, localization, and lateralization with bilateral cochlear implants," *J. Acoust. Soc. Am.*, vol. 113, no. 3, pp. 1617–1630, 2003.

[167] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 3, no. 3, pp. 247–251, Jul. 1993.

[168] N. J. Versfeld, J. M. Festen, and T. Houtgast, "Preference judgements of artificial processed and hearing-aid transduced speech," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1566–1578, 1999.

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

[170] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, vol. 1, no. 4, pp. 30–39, 1983.

[171] S. Voran, "A basic experiment on time-varying speech quality," in *Proc. International Conf. on Measurement of Speech and Audio Quality in Networks (MESAQIN)*, Prague, Czech Republic, Jun. 2005.

[172] M. Waltermann and A. R. andS.Moller, "Perceptual dimensions of wideband-transmitted speech," in *2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, Germany, September 2006, pp. 103–108.

[173] M. Waltermann, A. Raake, and S. Moeller, "Modeling of integral quality based on perceptual dimensions - a framework for a new instrumental speech-quality measure," *Voice Communication (SprachKommunikation), 2008 ITG Conference on*, pp. 1–4, oct. 2008.

[174] D. Wang, U. Kjems, M. Pedersen, J. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.*, vol. 125, pp. 2336–2347, 2009.

[175] J. Wen, N. D. Gaubitch, E. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

[176] J. Y. C. Wen and P. A. Naylor, "Semantic colouration space investigation: Controlled colouration in the bark-sone domain," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, oct. 2007, pp. 311–314.

[177] L. J. Williams, H. Abdi, R. French, and J. B. Orange, "A tutorial on multiblock discriminan correspondence analysis (mudica): A new method for analyzing discourse data from clinical populations," *Journal of Speech, Language and Hearing Research*, vol. 53, pp. 1372 – 1393, 2010.

[178] K. Worrall and R. Fellows, "Practical and affordable intelligibility testing for engineers and algorithm developers," in *Proc. AES Conf. on Audio Forensics*, Hillerod, Denmark, June 2010, pp. 194–201.

[179] W. Zha and W.-Y. Chan, "Voice quality assessment using classification trees," in *Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 1, November 2003, pp. 537 – 541.

[180] ——, "Objective speech quality measurement using statistical data mining," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1410 – 1424, 2005.

Appendix A

C-Qual Database

This appendix presents further details of the C-Qual database, first presented in Section 2.3.1 of this thesis. A description of the 44 degradation conditions and the condition averaged MOS obtained from the listening tests are presented in Table A.1.

Condition no.	Type of degradation	Description	Level of degradation	MOS
1	Additive noise	Car	-16 dB SNR	1.02
2	Additive noise	Car	-8 dB SNR	1.82
3	Additive noise	Car	0 dB SNR	3.03
4	Additive noise	Car	8 dB SNR	3.50
5	Additive noise	Car	16 dB SNR	4.05
6	Additive noise	Car	24 dB SNR	4.19
7	Additive noise	Car	32 dB SNR	4.52
8	Additive noise	Babble	-16 dB SNR	1.10
9	Additive noise	Babble	-8 dB SNR	1.15
10	Additive noise	Babble	0 dB SNR	2.03
11	Additive noise	Babble	8 dB SNR	3.20
12	Additive noise	Babble	16 dB SNR	3.78
13	Additive noise	Babble	24 dB SNR	4.18
14	Additive noise	Babble	32 dB SNR	4.11
15	Additive noise	Hum	-30 dB SNR	1.08
16	Additive noise	Hum	-20 dB SNR	1.52
17	Additive noise	Hum	-10 dB SNR	2.12
18	Additive noise	Hum	0 dB SNR	2.88
19	Additive noise	Hum	10 dB SNR	3.30
20	Additive noise	Hum	20 dB SNR	3.72
21	Additive noise	Hum	30 dB SNR	4.44
22	Clicks		160 clicks	2.07
23	Clicks		120 clicks	2.43
24	Clicks		40 clicks	2.98
25	Clicks		35 clicks	3.38
26	Clicks		20 clicks	3.97
27	Peak clipping		-8 dBFS	1.68
28	Peak clipping		-12 dBFS	1.88
29	Peak clipping		-16 dBFS	2.62
30	Peak clipping		-20 dBFS	3.22
31	Reverberation	MARDY RIR	1m	4.38
32	Reverberation	MARDY RIR	2m	4.02
33	Reverberation	MARDY RIR	3m	3.75
34	Reverberation	Office RIR	0.75m	3.89
35	Reverberation	Office RIR	1.85m	3.31
36	Coloration	low cut		3.60
37	Coloration	high boost		4.40
38	Coloration	tilt		3.69
39	MNRU		10	2.71
40	MNRU		15	2.24
41	MNRU		20	3.37
42	MNRU		25	3.68
43	MNRU		30	3.75
44	MNRU		50	3.78

Table A.1: The 44 degradation conditions in the C-Qual database, with corresponding condition averaged MOS.

Appendix B

Pitch estimation

This appendix presents histograms of pitch estimation error for the four algorithms described in Section 2.4.2.1.

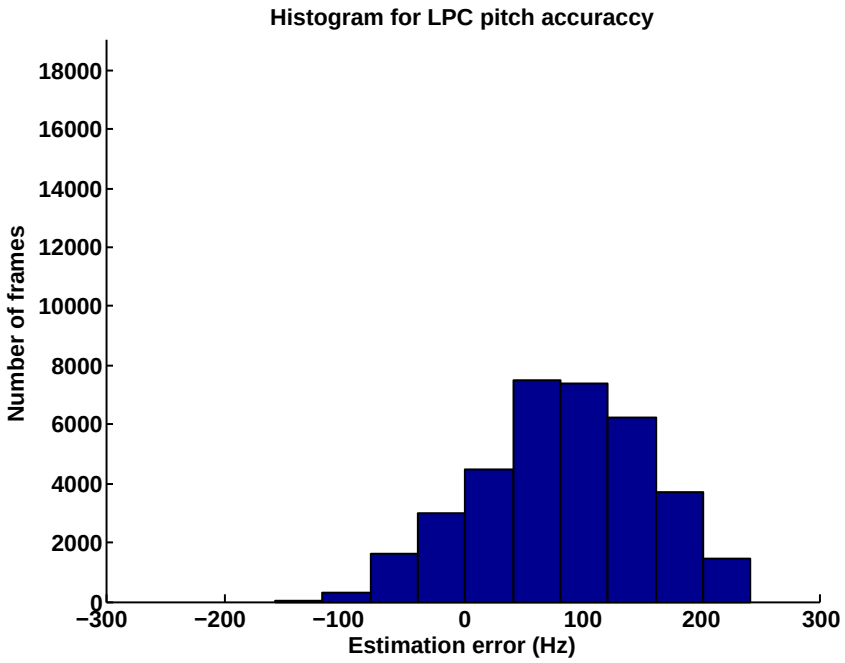

Figure B.3.1: Pitch estimation error histogram for the autocorrelation based pitch algorithm.

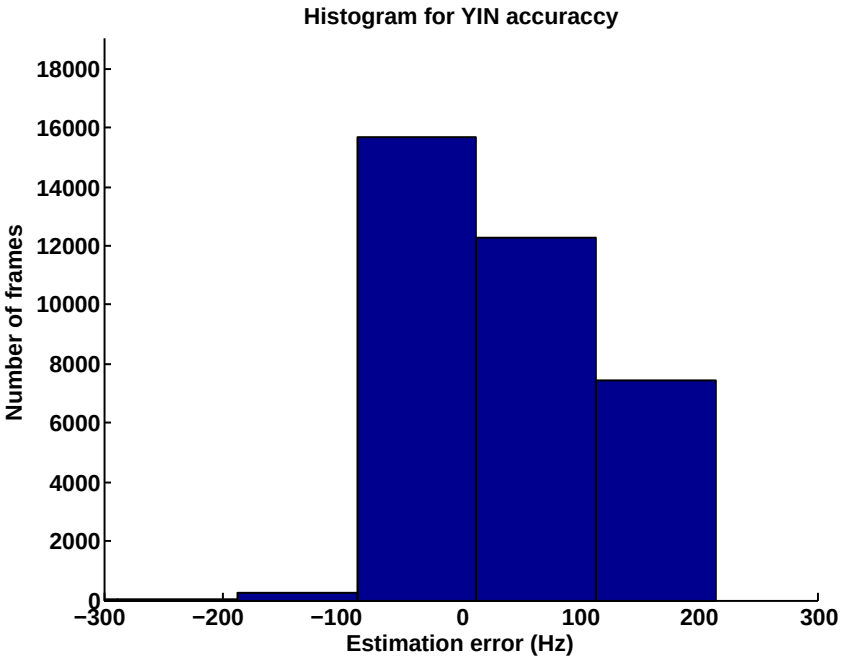

Figure B.3.2: Pitch estimation error histogram for the YIN algorithm.

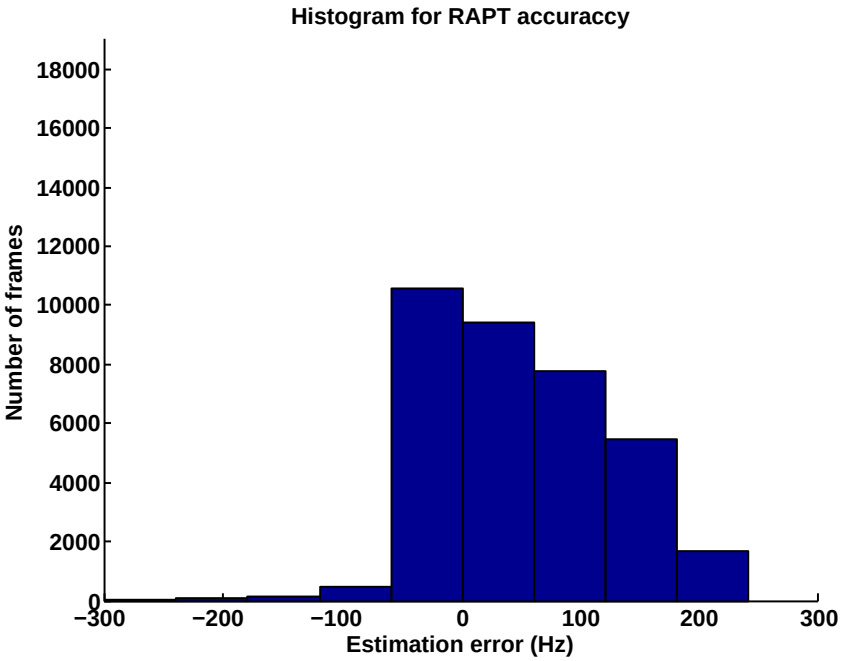

Figure B.3.3: Pitch estimation error histogram for the RAPT algorithm.

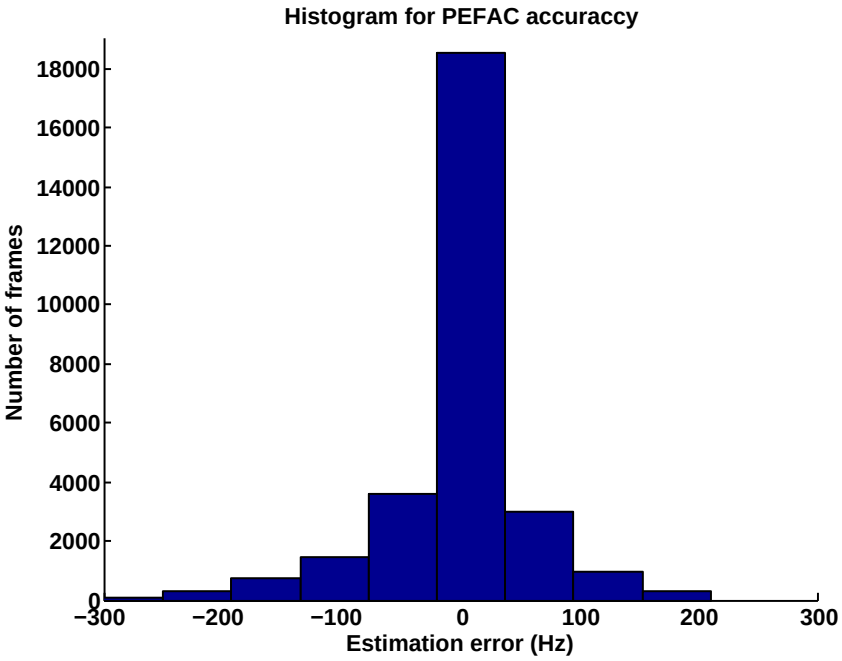

Figure B.3.4: Pitch estimation error histogram for the PEFAC algorithm.

Appendix C

TAXIT Database

This appendix presents further details of the TAXIT database that was used for the vocabulary labeling experiment in Section 6.4 of this thesis. Tables C.2 and C.3 present the 55 base conditions of the TAXIT database. The 55 base conditions are further processed by 4 CODEC arrangements as outlined in Table C.4.

Condition no.	Brick-wall filtering	Transition band (Hz)	Attenuation (dB)
1	Low pass	500:550	-60
2	Low pass	1000:1050	-60
3	Low pass	1500:1550	-60
4	Low pass	2000:2050	-60
5	Low pass	2500:2550	-60
6	Low pass	3000:3050	-60
8	High pass	500:550	-60
9	High pass	1000:1050	-60
10	High pass	1500:1550	-60
11	High pass	2000:2050	-60
12	High pass	2500:2550	-60
13	Band pass	Low: 1000:950 High: 3000:3050	-60
14	Band pass	Low: 1500:1450 High: 2500:2550	-60
15	Band pass	Low: 1500:1450 High: 2000:2050	-60
16	Band pass	Low: 1000:950 High: 1500:1550	-60
	Shelf filtering	**Centre Frequency (Hz)**	**Gain (dB)**
17	Low shelf	1000	-30
18	Low shelf	2000	-30
19	Low shelf	3000	-30
20	Low shelf	1000	30
21	Low shelf	2000	30
22	Low shelf	3000	30
23	High shelf	1000	-30
24	High shelf	2000	-30
25	High shelf	3000	-30
26	High shelf	1000	30
27	High shelf	2000	30
28	High shelf	3000	30
29	Clockwise tilt	2000, 2000	30, -30
30	Anti-clockwise tilt	2000, 2000	-30, 30

Table C.2: The base conditions of the TAXIT database.

Condition no.	Degradation	Description	SNR (dB)
31	Additive noise	Hum	0
32	Additive noise	Hum	-5
33	Additive noise	Hum	-10
34	Additive noise	Car	0
35	Additive noise	Car	-5
36	Additive noise	Car	-10
37	Additive noise	Babble	0
38	Additive noise	Babble	-5
39	Additive noise	Babble	-10
	Reverberation	**Distance (m)**	**DRR**
40	MARDY room [175]	1	19.10
41	MARDY room [175]	2	13.68
42	MARDY room [175]	3	10.82
43, 44	**Envelope fluctuations**		
	Temporal erasures	**Duration (ms)**	**Number**
45	Clicks	7	35
46	Clicks	7	70
47	Clicks	7	105
48	Drop outs	90	3
49	Drop outs	90	6
50	Drop outs	90	9
		Level (dBFS)	
51	Peak clipping	-20	
53	Peak clipping	-25	
53	Peak clipping	-30	
54	Clean, undegraded speech		
55	Clean, undegraded speech		

Table C.3: The 55 base conditions of the TAXIT database.

Condition no.	CODEC
1:55	PCM
56:110	GSM-FR
111:165	Transcoding: GSM-G711-GSM
166:220	MP3

Table C.4: The 220 degradation conditions of the TAXIT database. The 55 base conditions are processed by the four CODEC arrangements described above.