Non-intrusive Estimation of Acoustic Parameters from Degraded Speech

by
Mr. D. James Eaton
B. Sc. (Hons) B. A. (Hons) MIET

A Thesis submitted in fulfilment of the requirements for the
Doctor of Philosophy (PhD) Degree of Imperial College London, and the
Diploma of Imperial College London (DIC) by Research

Speech and Audio Processing Group
Communications and Signal Processing Division
Department of Electrical and Electronic Engineering
Imperial College London
May 2016

Declaration of originality

This work is the work of the author and that all other work is appropriately referenced.

Declaration of copyright

Abstract

ESTIMATION of the acoustic parameters Signal-to-Noise Ratio (SNR), Reverberation Time (T_{60}), Direct-to-Reverberant Ratio (DRR), and clipping level from degraded speech are open research questions. These parameters are important for determining speech quality and intelligibility, and they are widely applicable to speech enhancement and speech recognition systems. Whilst SNR, T_{60}, and DRR are useful priors for dereverberation schemes, indications of clipping and the clipping level are useful for signal restoration.

This thesis investigates how accurately and robustly to noise and, in the case of clipping detection, robustness to the coding and decoding process it is possible to estimate these parameters non-intrusively from degraded speech in real-time or near real-time, and introduces a range of novel algorithms. Alongside the algorithms, an international research challenge was staged for which a novel noisy reverberant speech corpus was developed to determine the state-of-the-art in T_{60} and DRR estimation.

In tests, the algorithms presented in this thesis were highly competitive, being able to estimate T_{60} with Pearson correlation coefficient, $\rho = 0.608$ and DRR with $\rho = 0.314$. Both algorithms achieved very low computational complexity with Real-Time Factors (RTFs) of 0.0164 and 0.0589 respectively. The clipping detection algorithms achieved F1 approximately 0.6 for Global System For Mobile Communications (GSM) 06.10 decoded speech, babble noise at 20 dB SNR clipping levels in the range -5 to -20 dBFS, and also produce an estimate of the original unclipped signal level.

Acknowledgment

I would like to thank Mr. Mike Brookes for providing assistance in performing the work described in Chapter 2 and 5. I would like to thank Dr. Nikolay D. Gaubitch for providing assistance in developing the algorithm described in Chapter 3, and Dr. Alastair H. Moore for providing assistance in the development of the algorithm described in Chapter 4, both of whom provided assistance in performing the ACE Challenge. I would like to thank Mr. Tom Neeld and Dr. David Shipworth for engaging with us in a fascinating collaboration with University College London.

I would like to thank the IEEE AASP Technical Committee for their support and endorsement of the ACE Challenge, the challenge participants for their research work on new algorithmic contributions, their results submissions, conference papers and posters. I would like to thank the IEEE WASPAA 2015 Committee for organizing the Acoustic Characterization of Environments (ACE) Challenge satellite workshop.

I would like to thank the postgraduate students and staff at Imperial College London and researchers at Delft University of Technology who provided their voices for the ACE corpus, and my colleagues at Imperial College London who have been a source of companionship and discussion in the laboratory.

I would like to thank the EPSRC, from which funding has made this research possible. I would like to thank Mr. Alan Worley, Clinical Scientist at Great Ormond Street Hospital, London, who has been a great source of encouragement and enthusiasm.

Finally I would like to thank my supervisor, Dr. Patrick A. Naylor, who has been an invaluable source of guidance, knowledge and inspiration throughout my studies.

Contents

List of figures

List of tables

List of notation

n	Sample	40
$x(n)$	Speech, sampled	40
$v(n)$	Noise, sampled	40
$y(n)$	Noisy speech, or noisy reverberant speech, sampled	40
$\mathbf{h}$	Room impulse response in discrete time	40
A_x	Active speech power determined using ITU-T P.56	45
ξ	*a priori* SNR	45
E	Expectation operator	45
$\hat{v}(n)$	Noise estimate, sampled	45
$\hat{v}^2(n)$	Noise power estimate, sampled	45
$\hat{\xi}$	Estimated *a priori* SNR	45
ϵ	Value to prevent estimation becoming undefined	45
$\hat{A}_x$	Estimated active speech power	45
ξ_s	Source SNR of unprocessed noise and speech sources	53
$\hat{\xi}_s$	Estimated source SNR	53
h_n	The n^{th} coefficient of the room impulse response, $\mathbf{h}$	61
n_0	Sample index at the arrival time of the direct path component	63
t	Time index	72
$h(t)$	Room impulse response, in continuous time	72
f	Frequency	72
H	Room impulse response spectrum	72
P	Power Spectral Density	72
λ	Room decay rate	73

λ_h	Room decay rate for a given RIR, h	72
$b(t)$	Stationary Gaussian noise process as a function of time	72
δ	Damping constant	72
σ^2	Variance	72
k	Frequency bin index	73
K	Number of frequency bins	73
$\sigma^2_{x^-}$	NSV of the speech signal	73
q	Mel frequency index	79
$\sigma^2_A(\lambda)$	Decay rate variance of T_{60} estimation mode A	79
$\sigma^2_B(\lambda)$	Decay rate variance of T_{60} estimation mode B	79
$\sigma^2_C(\lambda)$	Decay rate variance of T_{60} estimation mode C	79
$m(\cdot)$	Mapping function	79
f_s	Sample frequency	80
$x(t)$	Speech, continuous in time	93
$h_m(t)$	AIR at the m^{th} microphone	92
$v_m(t)$	Noise at the m^{th} microphone	93
$y_m(t)$	Noisy reverberant speech at the m^{th} microphone	92
$h_{m,d}(t)$	AIR of the direct sound at the m^{th} microphone	93
$h_{m,r}(t)$	AIR of the reverberant sound at the m^{th} microphone	93
$\bar{\gamma}$	Mean DRR	93
$\bar{\gamma}_m$	Mean DRR at the m^{th} microphone	93
$\mathbf{x}$	Vector of speech samples	94
$\mathbf{y}$	Noisy, or noisy reverberant speech, in vector notation	94
Γ	SRR	93
Γ_m	SRR at the m^{th} microphone	93
X	Unit plane wave spectrum	94
$\mathbf{x}$	Unit plane wave vector	94
Z	Delay-sum beamformer output spectrum	94
Ω	Look direction	94

G	Beamformer gain spectrum	95
$\mathbf{w}$	Beamformer complex weighting	94
W	Beamformer complex weighting spectrum	94
d	Beamformer direct signal	95
$\mathbf{d}$	Beamformer direct signal vector	95
r	Beamformer reverberant signal signal	95
$\mathbf{r}$	Beamformer reverberant signal vector	95
D	Beamformer direct signal spectrum	94
$S(j\omega)$	Speech spectrum	94
$H_{m,d}(j\omega)$	Beamformer direct signal AIR spectrum at the m^{th} microphone	94
$H_{m,r}(j\omega)$	Beamformer reverberant signal AIR spectrum at the m^{th} mic.	94
$X_m(j\omega, \Omega)$	Unit plane wave signal spect., look direction, for the m^{th} mic.	94
$\mathbf{x}(j\omega, \Omega)$	Unit plane wave signal spectrum, look direction	94
$D_m(j\omega)$	Beamformer direct signal spectrum at the m^{th} microphone	94
$R_m(j\omega)$	Beamformer reverberant signal spectrum at the m^{th} mic.	94
$Y_m(j\omega)$	Noisy reverberant speech spectrum at the m^{th} microphone	94
$V_m(j\omega)$	Noise spectrum at the m^{th} microphone	94
$R(j\omega)$	Beamformer reverberant signal	95
$Z_y(j\omega)$	Beamformer output spectrum	94
$Z_d(j\omega)$	Beamformer output spectrum direct path part	94
$Z_r(j\omega)$	Beamformer output spectrum reverberant path part	94
$Z_v(j\omega)$	Beamformer output spectrum noise part	94
$G^2(j\omega)$	Beamformer gain power spectrum	95
η	Clipping level	104
μ_+	Original signal level on positive side	103
μ_-	Original signal level on negative side	103
N	Length of the speech signal	109
$\hat{\mathbf{c}}$	Vector indicating which samples are estimated to be clipped	109
ϵ	Clipping tolerance	109

ψ_+	Threshold above which Atlas method considers a sample clipped 109
ψ_-	Threshold below which Atlas method considers a sample clipped 109
ζ	Normalisation factor applied to signal 113
u_o	Estimated original signal level 112
$\hat{u}_{o-}$	Negative estimated original signal level 114
$\hat{u}_{o+}$	Positive estimated original signal level 114
i_s	Histogram bin offset at the centre for least squares inner fit 114
i_m	Histogram bin offset at the outside for least squares inner fit ... 114
i_{0+}	First positive valued bin in the log log histogram 114
i_{min}	First negative valued bin in the log log histogram 114
i_{max}	Most positive valued bin in the log log histogram 114
i_{0-}	Most negative value bin in the log log histogram 114
L	Number of items in the least squares fit 114
a_+	Gradient of the positive top side 114
a_-	Gradient of the negative top side 114
b_+	Constant part of the line representing the positive side gradient 114
b_-	Constant part of the line representing the negative side gradient 114
d_+	Gradient of the positive side end slope 115
d_-	Gradient of the negative side end slope 115
e_+	Constant part of the positive side end slope line 115
e_-	Constant part of the negative side end slope line 115
f	Least squares function 114
ψ	ILAH minimum gradient 114
β	Maximum threshold upon which to perform clipping detection . 115
$\hat{u}_{p+}$	Positive amplitude intersect for perceptually decoded speech ... 116
$\hat{u}_{p-}$	Negative amplitude intersect for perceptually decoded speech .. 116
$\hat{u}_{p+min}$	Pos. inner bounds of the spread of values due to perc. codec .. 116
$\hat{u}_{p-min}$	Negative inner bounds of the spread of values 116
$\hat{P}$	Estimated perceptually decoded flag 116

χ_-	Histogram inner and outer gradient negative side split point	126
g_+	ILAH2 mapping function for positive excursions	127
g_-	ILAH2 mapping function for negative excursions	127
$\mathcal{TP}$	True positives - hits	128
$\mathcal{FP}$	False positives - false alarms	128
$\mathcal{TN}$	True negatives - correct rejections	128
$\mathcal{FN}$	False negatives - misses	128
$\mathcal{P}$	Positives (both true and false)	128
$\mathcal{N}$	Negatives (both true and false)	128
$\mathcal{FPR}$	False positive rate, or fall-out rate	128
ϵ_w	Atlas tolerance for clipped unencoded speech	129
ϵ_{oc}	Optimal Atlas tolerance for clean speech	130
$y_m(n)$	Noisy reverberant speech at the m^{th} microphone	141
$v_m(n)$	Noise at the m^{th} microphone	141
g	Inverse filter for AIR	146
$\mathbf{g}$	Inverse filter for AIR in vector notation	146

List of publications

Most of the author's own research can be found from chapter 2 of this thesis. Some of the research presented in this thesis can also be found in the following publications:

[1] James Eaton, Nikolay Gaubitch and Patrick A. Naylor

Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost.

ICASSP 2013

Abstract: Reverberation Time (T_{60}) is an important measure of the acoustic properties of a room. It can provide information about the acoustic environment, the intelligibility, and quality of speech recorded in the room, and help improve the performance of speech processing algorithms with reverberant speech. Where the acoustic impulse response of the room is not available, the T_{60} must be estimated non-intrusively from reverberant speech. State-of-the-art non-intrusive T_{60} estimators have been shown to be strongly biased in the presence of noise. We describe a novel T_{60} estimation algorithm based on spectral decay distributions that provides robustness to additive noise for a range of realistic noise types for signal-to-noise ratios in the range 0 to 35 dB and T_{60} between 200 and 950 ms. The proposed method also has much reduced computational cost.

This can be downloaded from:

`https://www.researchgate.net/publication/261230407_Noise-robust_reverberation_time_estimation_using_spectral_decay_distributions_with_reduced_computational_cost`

[2] James Eaton, Mike Brookes and Patrick A. Naylor

A Comparison of Non-Intrusive SNR Estimation Algorithms and the Use of Mapping

Functions.

EUSIPCO 2013

Abstract: We present a comparative evaluation of six methods for non-intrusive Signal-to-Noise Ratio (SNR) estimation for narrowband speech in noise. We demonstrate that the performance of all methods can be improved by applying a non-linear mapping function to their estimates of SNR. We have employed phrases built from the TI-MIT speech corpus (TIMIT) and noises from a broad range of sources including ITU-T P.501, NOISEX-92, and Soundjay. We compare the accuracy of the methods in estimating the SNR of both stationary and non-stationary noise and we conclude that with the mapping function, the best current methods can estimate the SNR to within approximately 3.5 dB for SNRs from −5 dB to 35 dB.

This can be downloaded from:

`https://www.researchgate.net/publication/274718080_A_Comparison_of_Non-Intrusive_SNR_Estimation_Algorithms_and_the_Use_of_Mapping_Functions`

[3] James Eaton and Patrick A. Naylor

Detection of clipping in coded speech signals.

EUSIPCO 2013

Abstract: In order to exploit the full dynamic range of communications and recording equipment, and to minimise the effects of noise and interference, input gain to a recording device is typically set as high as possible. This often leads to the signal exceeding the input limit of the equipment resulting in clipping. Communications devices typically rely on codecs such as Global System For Mobile Communications (GSM) 06.10 to compress voice signals into lower bitrates. Detecting clipping in a hard-clipped speech signal is straightforward due to the characteristic flattening of the peaks of the waveform, this is not the case for speech that has subsequently passed through a codec. We describe a novel clipping detection algorithm based on amplitude histogram analysis and least squares residuals which can estimate the clipped samples and the original signal level in speech even after

the clipped speech has been perceptually coded.

This can be downloaded from:

`https://www.researchgate.net/publication/274718060_Detection_of_Clipping_in_Coded_Speech_Signals`

[4] James Eaton and Patrick A. Naylor

Noise-robust detection of peak-clipping in decoded speech.

ICASSP 2014

Abstract: Clipping is a commonplace problem in voice telecommunications and detection of clipping is useful in a range of speech processing applications. We analyse and evaluate the performance of three previously presented algorithms for clipping detection in decoded speech in high levels of ambient noise. We identify a baseline method which is well known for clipping detection, determine experimentally the optimized operation parameter for the baseline approach, and use this in our experiments. Our results indicate that the new algorithms outperform the baseline except at extreme levels of clipping and negative signal-to-noise ratios.

This can be downloaded from:

`https://www.researchgate.net/publication/269295335_Noise-robust_detection_of_peak-clipping_in_decoded_speech`

[5] James Eaton, Alastair H. Moore, Patrick A. Naylor, and Jan Skoglund

Direct-to-reverberant ratio estimation using a null-steered beamformer.

ICASSP 2015

Abstract: Reverberation affects the quality and intelligibility of distant speech recorded in a room. Direct-to-Reverberant Ratio (DRR) is a useful measure for assessing the acoustic configuration and can be used to inform dereverberation algorithms. We describe a novel DRR estimation algorithm applicable where the signal was recorded with two or more microphones, such as mobile communications devices and laptops. The method uses a null-steered beamformer. In simulations the proposed method yields accurate DRR estimates to within ± 4 dB across a wide variety of room sizes, reverberation times and source-receiver distances. It is also

shown that the proposed method is more robust to background noise than a baseline approach. The best estimation accuracy is obtained in the region from -5 to 5 dB which is a relevant range for portable devices.

This can be downloaded from:

`https://www.researchgate.net/publication/275350843_Direct-to-reverberant_ratio_estimation_using_a_null-steered_beamformer`

[6] James Eaton, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor

The ACE Challenge - corpus description and performance evaluation.

WASPAA 2015

Abstract: Knowledge of the Direct-to-Reverberant Ratio (DRR) Reverberation Time (T_{60}) can be used to better perform speech and audio processing such as dereverberation. Traditionally, these parameters are computed from measured Acoustic Impulse Responses (AIRs). However, in many practical situations the AIR is not available and the parameters must be estimated non-intrusively directly from noisy speech or audio signals. The ACE Challenge is a competition to identify the most promising non-intrusive DRR and T_{60} estimation methods using real noisy reverberant speech. We the describe the ACE corpus comprising multi-channel AIRs, and multi-channel noise including ambient, fan and babble noise recorded in the same environment as the measured AIRs, along with the corresponding DRR and T_{60} measurements. The evaluation methodology is discussed and comparative results are shown. This can be downloaded from:

`https://www.researchgate.net/publication/280318558_The_ACE_Challenge_-_Corpus_Description_and_Performance_Evaluation`

[7] James Eaton, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor

Proceedings of the ACE Challenge Workshop - a satellite event of IEEE-WASPAA (2015)

The ACE Challenge Workshop - a satellite event of WASPAA 2015

Abstract: Several established parameters and metrics have been used to characterize

the acoustics of a room. The most important are the Direct-to-Reverberant Ratio (DRR), the Reverberation Time (T_{60}) and the reflection coefficient. The acoustic characteristics of a room based on such parameters can be used to predict the quality and intelligibility of speech signals in that room. Recently, several important methods in speech enhancement and speech recognition have been developed that show an increase in performance compared to the predecessors but do require knowledge of one or more fundamental acoustical parameters such as the T_{60}. Traditionally, these parameters have been estimated using carefully measured Acoustic Impulse Responses (AIRs). However, in most applications it is not practical or even possible to measure the acoustic impulse response. Consequently, there is increasing research activity in the estimation of such parameters directly from speech and audio signals. The aim of this challenge was to evaluate state-of-the-art algorithms for blind acoustic parameter estimation from speech and to promote the emerging area of research in this field. Participants evaluated their algorithms for T_{60} and DRR estimation against the 'ground truth' values provided with the data-sets and presented the results in a paper describing the method used. This can be downloaded from:

`http://arxiv.org/html/1510.00383`

[8] James Eaton and Patrick A. Naylor

Reverberation time estimation on the ACE corpus using the SDD method

The ACE Challenge Workshop - a satellite event of WASPAA 2015

Abstract: Reverberation Time (T_{60}) is an important measure for characterizing the properties of a room. The authors T_{60} estimation algorithm was previously tested on simulated data where the noise is artificially added to the speech after convolution with a impulse responses simulated using the image method. We test the algorithm on speech convolved with real recorded impulse responses and noise from the same rooms from the Acoustic Characterization of Environments (ACE) corpus and achieve results comparable results to those using simulated data. This can be downloaded from:

`http://arxiv.org/abs/1510.01193`

[9] James Eaton and Patrick A. Naylor

Direct-to-Reverberant ratio estimation on the ACE corpus using a Two-channel beamformer

The ACE Challenge Workshop - a satellite event of WASPAA 2015

Abstract: Direct-to-Reverberant Ratio (DRR) is an important measure for characterizing the properties of a room. The recently proposed DRR Estimation using a Null-Steered Beamformer (DENBE) algorithm was originally tested on simulated data where noise was artificially added to the speech after convolution with impulse responses simulated using the image-source method. This paper evaluates the performance of this algorithm on speech convolved with measured impulse responses and noise using the Acoustic Characterization of Environments (ACE) Evaluation corpus. The fullband DRR estimation performance of the DENBE algorithm exceeds that of the baselines in all Signal-to-Noise Ratios (SNRs) and noise types. In addition, estimation of the DRR in one third-octave ISO frequency bands is demonstrated. This can be downloaded from:

`http://arxiv.org/abs/1510.07546`

[10] Thomas Neeld, James Eaton, Patrick A. Naylor, and David Shipworth

A novel method of determining events in combination gas boilers: Assessing the feasibility of a single-point acoustic sensor

Building and Environment

Abstract: To assess the impact of interventions designed to reduce residential space heating demand, investigators must be armed with field-trial applicable techniques that accurately measure space heating energy use. This study assesses the feasibility of using a single-point acoustic sensor to detect gas consumption events in domestic Combination Gas-fired Boilers (C-GFBs). The investigation has shown, for the C-GFB investigated, the following events are discernible using a single-point acoustic sensor: demand type (hot water or central heating); boiler ignition; fan runtime and motor frequency; and pump runtime and motor frequency. A detection algorithm was developed to automatically identify demand type and burner ignition from sounds

made by the C-GFB with accuracies of 100% and 97% respectfully. Demand type was determined by training a naive Bayes classifier. Burner ignition was determined by detecting a low frequency (5-10 Hz) pressure pulse produced during ignition. The acoustic signatures of the fan and pump were identified manually along with their respective motor frequencies. Additional work is required to detect burner duration, deal with detection in the presence of increased noise and expand the range of boilers investigated. There are considerable implications resulting from the widespread use of such a device on improving understanding of space heating demand.
`http://www.sciencedirect.com/science/article/pii/S0360132316300361`

[11] James Eaton, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor
Estimation of room acoustics parameters: The ACE Challenge.
IEEE Trans. Audio, Speech, Lang. Process.
Abstract: Reverberation Time (T_{60}) and Direct-to-Reverberant Ratio (DRR) are important parameters which together can characterize sound captured by microphones in non-anechoic rooms. These parameters are important in speech processing applications such as speech recognition and dereverberation. The values of T_{60} and DRR can be estimated directly from the Acoustic Impulse Response (AIR) of the room. In practice, the AIR is not normally available, in which case these parameters must be estimated blindly from the observed speech in the microphone signal. The Acoustic Characterization of Environments (ACE) Challenge set out to determine the state-of-the-art in blind acoustic parameter estimation and also to stimulate research in this area. A summary of the ACE Challenge, and the corpus used in the challenge is presented together with an analysis of the results. Existing algorithms were submitted alongside novel contributions, the comparative results for which are presented in this paper. The challenge showed that T_{60} estimation is a mature field where analytical approaches dominate whilst DRR estimation is a less mature field where machine learning approaches are currently more successful.

[12] James Eaton and Patrick A. Naylor
ACE Challenge Results Technical Report

Speech and Audio Processing Group, Imperial College London

Abstract: This document provides the results of the tests of acoustic parameter estimation algorithms on the Acoustic Characterization of Environments (ACE) Challenge Evaluation dataset which were subsequently submitted and written up into papers for the Proceedings of the ACE Challenge. This document is supporting material for a forthcoming journal paper on the ACE Challenge which will provide further analysis of the results. This can be downloaded from:

`http://www.commsp.ee.ic.ac.uk/~sap/uploads/data/ACE/ACE_technical_report.pdf`

Statement of originality

The following aspects of the thesis are believed to be original contributions:

1. The use of a mapping function in conjunction with a noise estimator to improve the performance of global Signal-to-Noise Ratio (SNR) estimation for a speech utterance described in Chapter 2;

2. The method for estimating Reverberation Time (T_{60}) non-intrusively from noisy speech using (3.17) in Section 3.3.2, the SDD with Mel-spaced frequency bands and selective averaging (SDDSA) method;

3. The method of estimating Direct-to-Reverberant Ratio (DRR) non-intrusively using a null-steered beamformer described in Chapter 4;

4. The following methods for detecting clipping described in Chapter 5:

   (a) The time domain methods ILAH, and ILAH2, for estimating the original signal level and the clipping level of a speech signal that has passed through a perceptual codec as described in Sections 5.2 and 5.6;

   (b) The frequency domain method LSR, for detecting clipping in a speech signal that has passed through a perceptual codec as described in Section 5.3;

   (c) The hybrid time- and frequency domain methods LILAH, and CRACD, for detecting clipping and estimating the original signal level in a speech signal that has passed through a perceptual codec as described in Sections 5.4 and 5.5.

5. The Acoustic Characterization of Environments (ACE) Challenge international benchmarking competition organised under the auspices of the Institute of Electrical and Electronics Engineers (IEEE) Audio and Acoustic Signal Processing (AASP) Technical

Committee for the evaluation of acoustic parameter estimation algorithms. In particular, the novel corpus of Acoustic Impulse Response (AIR) measurements and matched noise recordings including multi-channel babble noise described in Chapter 6.

List of abbreviations

AASP Audio and Acoustic Signal Processing 139

ACC Accuracy 128

ACE Acoustic Characterization of Environments. A noisy reverberant speech corpus and IEEE challenge run by the SAP group at Imperial College 139

AIR Acoustic Impulse Response 137

AMR Adaptive Multi-Rate 103

ANC Adaptive Noise Canceller 64

ANU Australian National University 150

ASR Automatic Speech Recognition 180

ATR Advanced Telecommunications Research Institute International, Kyoto, Japan .. 68

AURORA Aurora Experimental Framework for the Evaluation of the Performance of Speech Recognition Systems under Noisy Conditions 49

BSS Blind Source Separation 64

C_{50} Clarity Index 180

C-GFB Combination Gas-fired Boiler 28

CASA Computational Auditory Scene Analysis 48

CELP Code-excited Linear Prediction 103

CRACD Codec-robust Automatic Clipping Detector 122

CSR-WSJ Continuous Speech Recognition Wall Street Journal Phase 1 database 68

CST Connected Speech Test speech corpus 67

DARPA Defense Advanced Research Projects Agency of the United States Dept. of Defense

DARPA-RMD DARPA 1000-Word Resource Management Database for Continuous Speech Recognition 50

DDR3 Double Data Rate Type Three 81

DENBE DRR Estimation using a Null-Steered Beamformer 91

Dev Development dataset of the ACE Challenge 149

DIRHA Distant-speech Interaction for Robust Home Applications multi-microphone multi-language acoustic speech corpus 139

DoA Direction-of-Arrival 151

DRR Direct-to-Reverberant Ratio 137

EER Equal Error Rate 121

EMIB Eigenmike Microphone Interface Box 143

EPSRC Engineering and Physical Sciences Research Council

Eval Evaluation dataset of the ACE Challenge 149

F1 F1 Score

FAU Friedrich-Alexander-Universität 150

FB Fullband 140

FDR Free-Decay Region 64

FFT Fast Fourier Transform 146

FISM Fast Image Source Method 69

FN False Negative

FNR False Negative Rate 129

FP False Positive

FPR False Positive Rate 128

GCC-PHAT Generalized Cross-Correlation with Phase Transformation method of estimating TDoA 97

GMM Gaussian Mixture Model. An approximation to an arbitrary probability density function that consists of a weighted sum of Gaussian distributions 109

GSM Global System For Mobile Communications 178

IBM Ideal Binary Mask

ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing . 179

IEEE Institute of Electrical and Electronics Engineers 31

ILAH Iterated Logarithm Amplitude Histogram clipping detection method 110

ILAH2 An improved Iterated Logarithm Amplitude Histogram clipping detection method 123

IPA International Phonetic Association 67

IRS Intermediate Reference System 53

ISO International Organization for Standardization 101

LILAH LSR-ILAH clipping detection method 121

LP Linear Prediction 66

LSR Least Squares Residuals clipping detection method 110

LTASS Long Term Average Speech Spectrum 118

LTI Linear Time Invariant 138

LUFS Loudness Units Full-Scale 142

MARDY Multichannel Acoustic Reverberation Database at York 68

MCS Multidimensional Colouration Space 68

MIT Massachusetts Institute of Technology 49

ML Maximum Likelihood 64

MMSE Minimum Mean Squared Error 168

MS Minimum Statistics 45

MSE Mean Square Error

MTF Modulation Transfer Function 65

NIRA Non-Intrusive Room Acoustics 152

NIST National Institute of Standards and Technology 47

NOISEX-92 Database to Study the Effect of Additive Noise on Speech Recognition Systems 70

NOIZEUS Noisy Speech Corpus for Evaluation of Speech Enhancement Algorithms . . . 48

NOSRMR Normalized Overall SRMR 158

NOSRMR Normalized Overall SRMR 158

NSRMR Normalised SRMR

NSRMR Normalised SRMR

NSV Negative-Side Variance . . . 152

OSRMR Overall SRMR . . . 158

PDF Probability Density Function . . . 48

PSD Power Spectral Density . . . 45

RIR Room Impulse Response

RMS Root Mean Square . . . 177

ROC Receiver Operating Characteristic . . . 128

RTF Real-Time Factor . . . 151

SB Subband . . . 147

SDD Spectral Decay Distributions . . . 152

SDDMSB SDD with Mel-spaced frequency bands . . . 165

SDDSA SDD with Mel-spaced frequency bands and selective averaging . . . 79

SDDSA-G SDDSA with Gerkmann noise estimator

SDDSA-H SDDSA with Hendriks noise estimator

SDRAM Synchronous Dynamic Random Access Memory . . . 81

SIREAC Simulation of REal Acoustics Software Tool . . . 68

SNR Signal-to-Noise Ratio . . . 140

SNT Subspace Noise Tracking algorithm . . . 50

SPEECON Speech Databases for Consumer Devices . . . 49

SPL Sound Pressure Level . . . 102

SPP Speech Presence Probability . . . 46

SRI SRI International. Formerly Standford Research Institute . . . 49

SRMR Speech-to-Reverberation Modulation Energy Ratio . . . 66

SRR Signal-to-Reverberation Ratio . . . 93

SSN Simultaneous Switching Noise . . . 53

STFT Short Time Fourier Transform . . . 168

STNR NIST's Speech-to-Noise Ratio Estimation Algorithm 46

T_{20} Reverberation Time to decay by 20 dB 65

T_{30} Reverberation Time to decay by 30 dB

T_{60} Reverberation Time to decay by 60 dB 137

TDoA Time-Difference-of-Arrival 97

TI Texas Instruments, Inc. 49

TIMIT TI-MIT speech corpus 48

TN True Negative

TP True Positive

UFRJ Federal University of Rio de Janeiro 150

VAD Voice Activity Detector 45

WADA Waveform Amplitude Distribution Analysis 47

WASPAA IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

WER Word Error Rate 44

WGN White Gaussian Noise 69

Chapter 1

Introduction

1.1 Motivation and aims

SOCIETY is dependent on communications devices that transmit the speech of a person in one location through a communications channel to another person or device in another location. Despite large increases in mobile data traffic, there is still a high demand for mobile voice services. In the UK in 2014 for example, the population made an average of 176 minutes of outgoing mobile telephone calls per person each month [13], and Ofcom[1] data shows that in 2013, average per-capita monthly mobile call minutes increased in most of its comparator countries [14].

Speech signals received from the microphones of a communications device are usually degraded. These degradations are typically either additive, convolutive, or non-linear. Examples of additive degradations are ambient noise, and sensor noise from the microphones in the device. Convolutive degradations emanate from reverberation in the acoustic environment where the microphone is situated. The talker may at times be too close to the microphones of the device, or be speaking at too high a level for the device. This may result in the speech signal exceeding the level the device can tolerate, leading to amplitude clipping of the speech signal, a non-linear degradation. The channel may add further electrical interference (additive), echoes (convolutive), and distortion (non-linear) including codec distortion.

[1] Independent regulator and competition authority for the UK communications industries

Users of communications devices increasingly expect to hear more intelligible and better quality speech. In law enforcement, speech may be recorded from a single microphone a long distance from the speaker in environments with very high levels of noise, frequently with negative Signal-to-Noise Ratios (SNRs). Extracting intelligible speech as well as reliable information about the acoustic scene from such sources is crucial for bringing about prosecutions. Speech degradation has also been shown to increase cognitive load [15] therefore reducing the listener's capacity to perform other tasks at the same time. It is therefore desirable and in some cases necessary to be able to enhance or restore speech before it reaches the recipient to increase quality, intelligibility and comfort.

Speech enhancement algorithms typically require knowledge of the acoustic environment of the degraded speech. Non-intrusive estimates of the characteristics of the channel can be used to improve the performance of speech enhancement algorithms provided that the estimates are sufficiently accurate. Where information about the channel such as the Acoustic Impulse Response (AIR) is not available, the characteristics of the channel must be estimated blindly, or 'non-intrusively'. Enhancement and restoration schemes such as dereverberation may require knowledge of the Reverberation Time (T_{60}) [16] and Direct-to-Reverberant Ratio (DRR) [17]. Clipping restoration algorithms typically require knowledge of the clipped samples [18], and the parameter estimators themselves may need prior information such as SNR [1]. Applications such as speech-to-text rely on speech recognition, where knowledge of the reverberation time and SNR can facilitate selection of a training database recorded in a similar environment, thus reducing errors and increasing the speed of recognition.

Recently, manufacturers have begun to incorporate auxiliary microphones into steadily more powerful mobile communications devices incorporating multi-channel dereverberation and denoising which often provide better performance than single channel algorithms. The requirement for single channel speech enhancement will continue for some time however in areas such as hearing aids, where computational performance is constrained by the size of the device and the power consumption.

This thesis aims to determine how accurately acoustic parameters can be estimated non-intrusively from single- and multi-channel speech with additive, convolutive and non-linear degradations, and with the objective of better enabling speech enhancement, speech

recognition, and signal restoration in real-time or near real-time.

1.2 Summary of the thesis

In Chapter 2, a range of established and state-of-the-art noise estimators are compared with the objective of finding the best performing algorithm for non-intrusively estimating the overall SNR of a speech signal degraded by additive noise as shown in Fig. 1.1.

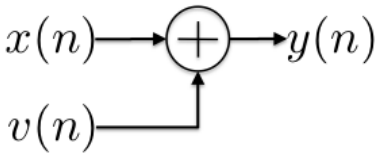

Figure 1.1: Signal flow diagram for Chapter 2, where $x(n)$ is the clean speech signal, $v(n)$ is the additive noise, and $y(n)$ is the speech signal degraded by additive noise

In order to improve upon the state-of-the-art, a mapping function trained on a speech corpus is introduced and applied to state-of-the-art algorithms to obtain the highest levels of performance. Phrases built from the TIMIT [19] speech corpus and noises from a broad range of sources including ITU-T P.501, NOISEX-92 [20], and Soundjay [21] were employed. The accuracy of the methods in estimating the SNR of both stationary and non-stationary noise was compared. It is found that with the mapping function, the best current method can estimate the SNR with a spread of 7.3 dB for SNRs from −5 dB to 35 dB for 95% of the results.

In Chapter 3, an algorithm for the estimation of acoustic parameters in speech with both additive and convolutive degradations as shown in Fig. 1.2 is presented. An

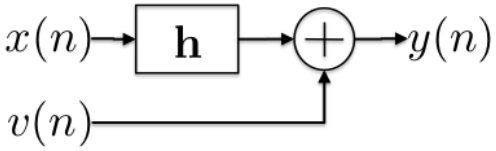

Figure 1.2: Signal flow diagram for Chapter 3, where $x(n)$ is the clean speech signal, h is the acoustic channel for the speech signal, $v(n)$ is the additive noise, and $y(n)$ is the speech signal degraded by additive noise

established method for estimating T_{60} that in common with other algorithms shows a strong bias in the presence of additive noise is used as the basis for a novel improved T_{60}

estimator robust to noise and with lower computational complexity. The algorithm requires an estimate of the SNR. In combination with noise estimators reviewed in Chapter 2, it is shown that the T_{60} estimation accuracy using the state-of-the-art in SNR estimation is almost as good as when the SNR is known *a priori*. The algorithm is compared with existing estimators using simulated data and it is shown that T_{60} can be estimated to with a maximum Root Mean Square (RMS) error of 120 ms for SNRs of 5 dB and above.

Continuing with the theme of additive and convolutive speech degradations, in Chapter 4, a novel low-complexity method for estimating DRR from two-channel speech using a null-steered beamformer is presented. The algorithm is compared with existing estimators using simulated data. The proposed method yields DRR estimates with a standard deviation of 4 dB across a wide variety of room sizes, reverberation times and source-receiver distances. It is also shown that the proposed method is more robust to background noise than a baseline approach. The best estimation accuracy is obtained in the region from −5 to 5 dB which is a relevant range for portable devices.

In Chapter 5, parameter estimation for speech with non-linear and additive degradations, as shown Fig. 1.3, is investigated. Several novel time- and frequency domain

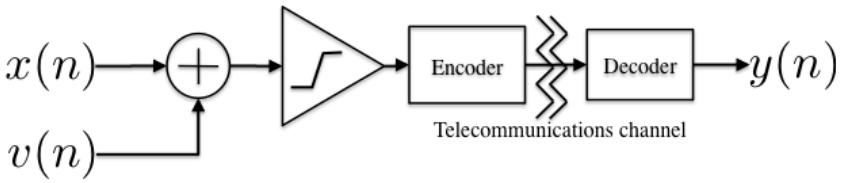

Figure 1.3: Signal flow diagram for Chapter 6, where $x(n)$ is the clean speech signal, $v(n)$ is the additive noise, and $y(n)$ is the speech signal degraded by additive noise, non-linear distortion and encoding and decoding

methods for detecting clipping are presented that are robust to the effects of encoding and decoding on speech, a research problem not previously attempted to be solved in the literature. The algorithms are evaluated in both unencoded and decoded speech in different levels of noise in comparison with a baseline optimized for operation in these conditions. The results indicate that the performance of the proposed algorithms outperforms the optimized baseline in Global System For Mobile Communications (GSM) 06.10 and Adaptive Multi-Rate (AMR)-decoded noisy speech with signal levels up to three times overdriven and SNRs down to 0 dB. The proposed methods also provide an estimate of the original signal amplitude before it was clipped.

In Chapter 6, the Acoustic Characterization of Environments (ACE) Challenge is presented. This was an international research challenge run by the author to determine the state-of-the-art in T_{60} and DRR estimation, and to stimulate new research. In the first part of the challenge a novel noisy reverberant speech corpus, the ACE Corpus, was developed to provide test signals with convolutive degradations for speech signals, and also additive noise with convolutive degradations from the same acoustic environment, as shown in Fig. 1.4. The ACE Corpus was designed, recorded, analyzed and distributed as part of this

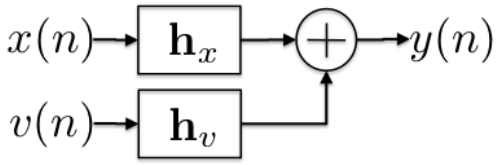

Figure 1.4: Signal flow diagram for Chapter 6, where $x(n)$ is the clean speech signal, $\mathbf{h}_x$ is the acoustic channel for the speech, $\mathbf{h}_v$ is the acoustic channel for the noise, $v(n)$ is the additive noise, and $y(n)$ is the speech signal degraded by additive noise. This signal flow represents a single microphone channel. For multi-channel signals, this flow is duplicated for each additional channel used.

research project. The ACE Corpus was then used as the basis for an international research challenge to determine the state-of-the-art in T_{60} and DRR estimation. The estimation performance of the author's T_{60} and DRR estimators is also compared using the real data from the ACE Corpus.

In Chapter 7, conclusions are drawn regarding what this work has achieved, and whether estimating acoustic parameters from degraded speech in single- and multi-channel microphone configurations can be performed with sufficient accuracy to enable speech enhancement and recognition.

Chapter 2

Non-intrusive Signal-to-Noise Ratio estimation from degraded speech

2.1 Introduction

NOISE is an almost universally present problem in speech processing systems [22]. Knowledge of the Signal-to-Noise Ratio (SNR) or the noise level is a fundamental building block in robust speech detection and enhancement algorithms as a means of identifying, extracting and enhancing speech from noisy sources. The level and characteristics of the noise are not typically known *a priori* and if required they must therefore be estimated.

Noise estimation is employed in a wide range of applications including automotive hands-free kits, hearing aids, teleconferencing terminals, set-top boxes, home automation systems, mobile telephones, smartphones, and personal digital assistants. It is employed in a range of industries including the military, law enforcement, and telecommunications. Knowledge of the SNR can also be useful when estimating other acoustic parameters such as Reverberation Time (T_{60}) and Direct-to-Reverberant Ratio (DRR), as will be shown in Chapter 3 [1], and Chapter 4 [5]. A further application of SNR estimation is audio diarization: the process of annotating an input audio channel with information that helps to classify (possibly overlapping) temporal regions of the signal to specific sources. Diarization

has utility in Automatic Speech Recognition (ASR) of multi-talker speech signals [23] and in searching and indexing audio archives [24].

Motivated towards the application of surveillance in law enforcement for which negative SNRs are common, this chapter is concerned with estimating the Fullband (FB) SNR of an entire speech utterance non-intrusively from noisy speech. This is also known as the global SNR.

Although ASR is an important application of speech enhancement, testing of Word Error Rate (WER) improvements for example are beyond the remit of this thesis. This Chapter only assesses work done in the research community or where the method was publicly available, and whilst commercial organizations may have good noise reduction systems, these are beyond the scope of this work.

2.1.1 Signal model

The SNR for noisy speech is defined as the ratio of the speech active level to the estimated mean noise power. The speech active level is defined as the average power of speech present in a signal and a method for estimating this is defined in ITU-T P.56 [25]. 'Non-intrusive' algorithms are defined in this context to mean algorithms that can operate without knowledge of the speech and the noise *a priori*, nor require them to be available as separate sources. For non-intrusive SNR estimation to be useful for acoustic parameter estimation, speech enhancement and audio diarization applications, it must be able to identify speech in high levels (such as 0 dB SNR) of both stationary and non-stationary noise over several utterances.

Let a noisy speech signal be expressed as

$$y(n) = x(n) + v(n) \tag{2.1}$$

where $x(n)$ is the clean speech signal including speech pauses, and $v(n)$ is uncorrelated noise. Let A_x be the estimated speech active level as defined in ITU-T P.56 [25] without pauses. Since the noise is uncorrelated, A_x can be considered to be equal to $E\{|x(n)|^2\}$

where the pauses are set to zero such that

$$E\{|y(n)|^2\} = A_x + E\{|v(n)|^2\}, \tag{2.2}$$

where $E\{\cdot\}$ is the expectation operator. The *a priori* SNR, ξ, is given by

$$\xi = \frac{A_x}{E\{|v(n)|^2\}} \tag{2.3}$$

In order to estimate the *a priori* SNR, either the mean noise power (or noise Power Spectral Density (PSD)) given noisy speech can be estimated, or else an estimate of the mean clean speech power given the noisy speech can be estimated. Assuming an estimate of the noise power $E\{|\hat{v}(n)|^2\}$, has been obtained, from (2.2) and (2.3), the estimated SNR, $\hat{\xi}$, can be computed as

$$\hat{\xi} = \frac{\hat{A}_x}{E\{|\hat{v}(n)\}|^2\}} = \frac{E\{|y(n)|^2\}}{E\{|\hat{v}(n)|^2\}} - 1. \tag{2.4}$$

or when only the noisy speech power and an estimate of the clean speech power is known:

$$\hat{\xi} = \frac{1}{\frac{E\{|y(n)|^2\}}{\hat{A}_x} - 1.} \tag{2.5}$$

The estimated SNR, $\hat{\xi}$, can therefore be determined either from the noisy speech power and the estimated noise PSD, or the noisy speech power the estimated clean speech active level.

2.1.2 Methods for non-intrusive SNR and noise estimation in the literature

Approaches to non-intrusive SNR and noise estimation in the literature can be considered in several classes: The first class uses a Voice Activity Detector (VAD) [26] to identify when speech is present in the signal by determining where the signal exceeds a predefined threshold, giving the noisy speech power, and then sampling the noise during the pauses to estimate the noise level. The SNR is then given by (2.4).

A second class of algorithms first proposed in [27] tracks the minimum of the magnitude spectrum over a moving window which provides biased estimates of the noise level (Minimum Statistics (MS)). These algorithms assume that within a sufficiently narrow frequency band, the speech energy will regularly fall to zero. Bias compensation has been

employed in several automatic noise reduction algorithms including Martin [28] and latterly Hendriks [29]. A recent innovation proposed in [30] dispenses altogether with the bias compensation and safety net in [29] and uses Speech Presence Probability (SPP) with fixed priors to determine how much of the current noisy speech or previous noise estimate to include in the noise estimate for the current frame.

A third class of algorithms estimates the noise and speech simultaneously using for example analysis of histograms such as [31, 32] or binary masks [33].

A fourth class of algorithms attempts to model the noise commonly assuming a Gaussian process and a slowly changing power spectrum such as [34].

A broad cross-section of algorithms that can be used to estimate SNR is described below. There have been many incremental improvements in the field, so this survey represents only the more significant developments.

ITU-T P.56

Method B of the ITU-T P.56 [25] standard determines a speech activity factor representing the proportion of time the signal is considered to be active speech. The algorithm works by first exponentially smoothing the speech twice, then comparing the smoothed speech level with a set of thresholds generated from the input signal. After analysing whether the speech exceeds each threshold, the threshold is found for which the active speech power is 15.9 dB above the threshold level power. The algorithm returns the speech activity level and activity factor. The SNR is obtained from (2.5) using the estimated speech power obtained by multiplying the active speech power with the activity factor, and the noisy speech power. This algorithm is not intended for use with SNRs below approximately 20 dB since noise occurring during the noisy speech is included in the active speech power estimate [35, 36]. A widely used implementation of ITU-T P.56 is *activlev.m* [37].

NIST STNR

The NIST's Speech-to-Noise Ratio (STNR) estimation algorithm uses a time-domain energy histogram based approach [31]. The algorithm returns the estimated SNR directly. The energy histogram is computed from the entire speech file using a 20 ms window with a 50%

overlap. The technique is based on the assumption that the mean speech energy and noise energy are different and therefore occupy different regions in the histogram. Using fitting techniques, these regions are identified and the corresponding noise level subtracted from the corresponding speech level to estimate the SNR.

Minimum statistics

Martin *et al.* [27] proposed an algorithm to estimate the PSD of non-stationary noise based on the assumption that in a given frequency band, the level of the speech will frequently drop to the level of the noise. The minimum statistics-based algorithm for estimating noise eliminates the problem of using a VAD as in previous noise PSD estimators [27]. The authors demonstrated that the performance is similar to previous algorithms in stationary noise but improved in non-stationary noise. Martin *et al.* [38], and Mauler *et al.* [39] proposed further improvements to the algorithm proposed in [27] and claim to have developed an unbiased noise estimator. A widely used implementation of [28] is *estnoisem* [37], which replaces Table 3 by the updated Table 5 from [38]. The slight improvement reported in [39] is not included in this implementation. The SNR is obtained from (2.4) using the estimated noise power and the noisy speech power.

WADA

The SNR estimation algorithm proposed by Kim *et al.* [32] is based on the observation that the value of the parameter α in a Gamma distribution characterises the SNR in the amplitude distribution of noisy speech. The authors claim that the Waveform Amplitude Distribution Analysis (WADA) algorithm has significantly less bias and variability with noise type in comparison with National Institute of Standards and Technology (NIST)'s STNR algorithm along with low computational complexity. The algorithm returns the estimated SNR directly.

MMSE-based noise PSD tracker with low complexity

Hendriks *et al.* [29] proposed a Minimum Mean Squared Error (MMSE)-based noise PSD tracker with low complexity. This algorithm improves on previous attempts to correct

the bias empirically by finding an analytical relationship between the biased and unbiased estimates by making assumptions about the Probability Density Functions (PDFs) of speech and noise. The MMSE is estimated and then the bias compensation applied to provide the estimate of the noise PSD. A safety net as proposed in [40] is applied to overcome locking of the algorithm should the noise level abruptly change between samples, and the results are then smoothed using recursive averaging. The authors claim similar performance to other state-of-the-art algorithms but with much reduced computational complexity. The SNR is obtained from (2.4) using the estimated noise power and the noisy speech power.

Noise PSD tracker with low complexity and low tracking delay

Gerkmann *et al.* [30] proposed a PSD tracker with low complexity and low tracking delay by interpreting an MMSE optimal estimator as a VAD-based noise power estimator, and then replacing the VAD with a soft SPP detector with fixed priors. The authors claim similar performance to other state-of-the-art algorithms but with further reduced computational complexity. The SNR is obtained from (2.4) using the estimated noise power and the noisy speech power. An efficient Matlab implementation of this algorithm is available as *estnoiseg.m* [37] providing a further reduction in computational complexity.

Non-real-time algorithms

Other algorithms such as those involving Computational Auditory Scene Analysis (CASA) [33] are also presented in the literature. However, these algorithms do not currently offer the potential for real-time operation and so they have not been considered in further detail in this thesis.

2.1.3 Corpora typically used to evaluate SNR estimation algorithms

Table 2.1 provides a survey of recent studies with the speech and noise corpora used. The most popular corpora were the TI-MIT speech corpus (TIMIT), NOISEX-92 and the Noisy Speech Corpus for Evaluation of Speech Enhancement Algorithms (NOIZEUS).

The TIMIT speech corpus

Whilst many corpora have been compiled for research into speech enhancement, a survey of the literature shows that over half the studies mentioned in this review used the TIMIT speech corpus. The TIMIT corpus of read speech was designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems [19]. The TIMIT corpus contains recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences each approximately 3 seconds long. The corpus design was a joint effort between the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI). Both test and training subsets, balanced for phonetic and dialectal coverage, are specified.

The NOISEX-92 noise corpus

Where noise sources were specified in the studies, the NOISEX-92 corpus was used more than any other noise corpus. The NOISEX-92 noise corpus was created as part of an experiment to study the effect of additive noise on speech recognition systems and provides a range of stationary and non-stationary noises [20].

The NOIZEUS noisy speech speech

A further widely used corpus is NOIZEUS, which uses noises from the Aurora Experimental Framework for the Evaluation of the Performance of Speech Recognition Systems under Noisy Conditions (AURORA) corpus [41]. The corpus provides speech mixed with noise at different SNRs.

2.1.4 Comparative evaluations in the literature

Recent studies of SNR and noise estimators are listed in Table 2.1. These studies are diverse and cover a wide range of different algorithms, speech and noise corpora, and test SNRs. Of these studies, Meyer *et al.* [42] compared algorithms proposed prior to 1997. Cohen *et al.* [43] used twenty utterances and noise between -5 and 10 dB using algorithms proposed prior to 2004. Vondrasek *et al.* [44] employed test data from CAR2ECS [45] and Speech Databases for Consumer Devices (SPEECON) [46]. Hu *et al.* [47] conducted a subjective

evaluation of 13 algorithms using the NOIZEUS [47] corpus. Beritelli *et al.* [48] considered vowel sounds only and evaluated non-automatic and semi-automatic methods using the TIMIT speech corpus. Kim *et al.* [32] compared the WADA algorithm with NIST STNR using the DARPA 1000-Word Resource Management Database (DARPA-RMD) [49]. Hendriks *et al.* [29] compared five algorithms in non-stationary noise in the range 0 to 15 dB. Taghia *et al.* [50] compared a broad range of estimators for a limited range of noises including the Subspace Noise Tracking (SNT) algorithm [51] over the range −5 to 20 dB SNR. However, Taghia did not evaluate Gerkmann *et al.* [30] which promises similar accuracy to Hendriks *et al.* [29] with much reduced computational complexity.

Study	Method	Speech source	Noise source
Meyer, 1997 [52]	Various	Not stated	Not stated
Martin, 2001 [28]	MMSE	Not stated	Not stated
Cohen, 2004 [43]	Non-causal estimator	TIMIT [19]	NOISEX-92 [20]
Vondrasek, 2005 [44]	VAD	CAR2ECS [45]	SPEECON [46]
Hu, 2006 [47]	Various	NOIZEUS [47]	NOIZEUS [47]
Kim, 2008 [32]	WADA	DARPA-RMD [49]	WGN, Speech, Music
Hendriks, 2010 [29]	MMSE	Not stated	Not stated
Beritelli, 2010 [48]	VAD	TIMIT [19]	Not stated
Taghia, 2011 [50]	Various	TIMIT [19]	Sound-Ideas [53]
Gerkmann, 2012 [30]	SPP	TIMIT [19]	Not stated
Narayanan, 2012 [33]	IBM/CASA	TIMIT [19]	NOISEX-92 [20]

Table 2.1: Speech and noise corpora by algorithm

Considering the speech and noise corpora used to evaluate noise and SNR estimation algorithms, and previous comparative evaluations, a comparison comparing [29] and [30] with established algorithms over a wide range of noise types and levels is relevant. An evaluation will have most impact and be able to be compared with the largest number of other recent studies if it uses the TIMIT speech corpus and the NOISEX-92 noise corpus supplemented with non-stationary noise from other sources since the number of non-stationary noises in NOISEX-92 is limited. Also, since estimating SNR at very high levels of noise or very low levels of noise represents the greatest challenge, a test at lower noise levels than 20 dB SNR as used in [50], from −5 to 35 dB SNR, will give a more comprehensive evaluation of the algorithms' performance. A further opportunity which has not been explored in the literature is to improve the estimation accuracy of a given algorithm using

a polynomial mapping function determined by analysing a speech corpus such as TIMIT [19] to map the SNR estimates onto the ground truth.

2.1.5 Aims and organisation of this chapter

The aim of this chapter is to present a comparative evaluation of SNR estimation algorithms for non-intrusive SNR estimation for narrowband speech in stationary and non-stationary noise. In order to maximise the performance of each algorithm, the use of mapping functions is investigated. Speech pauses will be included in the test utterances such as may occur in natural speech but are not always included in previously used test datasets.

Furthermore, included in this study is a wider range of SNR levels than has previously been investigated. The remainder of this chapter is organised as follows: In Section 2.2, the algorithms to be evaluated are described. In Section 2.3, the approach to evaluating the performance of each algorithm, the data used, and the training and test procedures employed are described. In Section 2.4 the experimental results are reviewed and the performance of the algorithms tested, and in Section 2.5 conclusions are drawn based on the results.

2.2 Non-intrusive SNR estimation algorithms

In order to make a broad meaningful comparison, both established and state-of-the-art algorithms using a range of different techniques were chosen. The algorithms chosen were:

1. ACT, an implementation [37] of ITU-T P.56 [25] that estimates speech activity level and duty cycle. The algorithm returns the speech activity level and activity factor. The SNR is obtained from (2.5) using the estimated speech power obtained by multiplying the active speech power with the activity factor, and the noisy speech power. This algorithm is not intended for use with low SNRs since noise occurring during the noisy speech is included in the active speech power estimate.

2. STN, the National Institute of Standards and Technology's stnr SNR estimation algorithm [31]. The algorithm returns the estimated SNR directly.

3. KSM, an implementation of Kim and Stern's SNR estimation algorithm [32] in Matlab which interpolates the values in the look up table providing an improvement in estimation accuracy. The algorithm returns the estimated SNR directly.

4. EST, an implementation of [28], with Table 3 replaced by the updated Table 5 from [38]. A slight improvement was reported by Mauler and Martin [39] but this is not included in the implementation [37]. The SNR is obtained from (2.4) using the estimated noise power and the noisy speech power.

5. ESG, an implementation of Gerkmann's noise PSD tracker [30] providing a reduction in computational complexity [37]. The SNR is obtained from (2.4) using the estimated noise power and the noisy speech power.

6. HEN, R. Hendrik's MMSE-based noise PSD tracker with low complexity [29] Matlab code. The SNR is obtained from (2.4) using the estimated noise power and the noisy speech power.

The use of a novel polynomial mapping function between the algorithms' outputs and the true SNR determined for each algorithm is proposed. This mapping was determined using the training corpus of the TIMIT [19] speech corpus, and stationary and non-stationary training noises were obtained from the ITU-T P.501 [54] monaural noise sequences. Coefficients for a fifth order polynomial fit in a least squares sense were computed over the range $-5\,\mathrm{dB}$ to $35\,\mathrm{dB}$ in $5\,\mathrm{dB}$ steps from the estimated SNR from each algorithm to minimise the maximum dB error against the true SNR at these points. A fifth order polynomial was chosen because it was found to give better improvements in accuracy than a third order polynomial. However, the improvement with a seventh order polynomial fit was not significant, and with higher orders, there is the risk of overfitting and unpredictable behaviour outside of the range of SNRs used in training.

2.3 Performance evaluation

Phrases from the TIMIT [19] training and test corpora simulating realistic clean speech were assembled by incorporating pauses into each utterance. Composite utterances comprising three concatenated TIMIT files separated by one second pauses were constructed totalling

448 speech files for the test dataset and 1152 speech files for the training dataset. TIMIT sentences SA1 and SA2 were excluded since these are repeated for each speaker. Stationary test dataset noises were obtained from the NOISEX-92 noise corpus [20]. Non-stationary test dataset noises were obtained from Soundjay [21] and modulated white noise and Simultaneous Switching Noise (SSN) provided by R. Hendriks as used in [29].

To obtain a noisy speech signal at the test SNR, ξ, each noisy speech signal, $y(n)$, was constructed by mixing the clean speech signal, $x(n)$, with appropriate noise, $v(n)$, as

$$y(n) = x(n) + \frac{v(n)\sqrt{\xi_s}}{\sqrt{\xi}} \tag{2.6}$$

where ξ_s is power ratio of the unprocessed speech and noise signals. In practice this must be estimated as follows

$$\hat{\xi}_s = \frac{\hat{A}_x}{E\{|v(n)|^2\}}. \tag{2.7}$$

Table 2.2 lists the noises used in the training and test datasets respectively.

To simulate narrowband telephony channel conditions, the Intermediate Reference System (IRS) [55] weighting was applied to the noisy speech signals. This frequency response of the IRS filter is shown in Fig. 2.1. Clean speech phrases and noises truncated to the

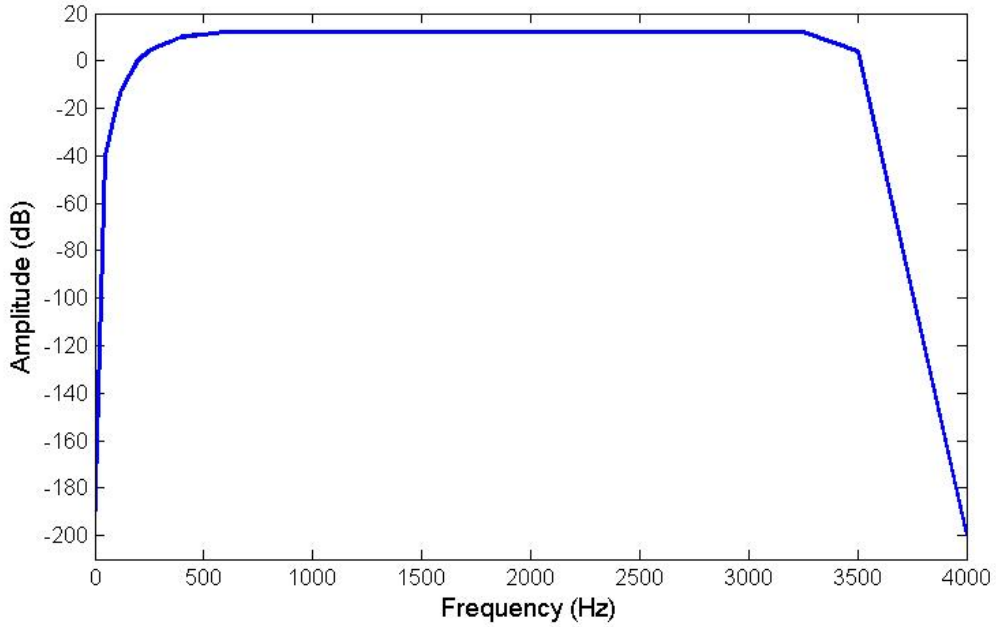

Figure 2.1: IRS filter characteristic from NOIZEUS

Training dataset	
Stationary noises	Non-stationary noises
Cafeteria (P.501)	Bus interior (Soundjay)
Office (P.501)	Car interior 2 (P.501)
Railway station (P.501)	Paris Metro (P.501)
Restaurant (P.501)	Street (P.501)
Car interior 1 (P.501)	

Test dataset	
Stationary noises	Non-stationary noises
Babble (NOISEX-92)	highway (soundjay)
Factory1 (NOISEX-92)	intersection (soundjay)
Factory2 (NOISEX-92)	Modulated SSN noise
Volvo car interior (NOISEX-92)	Modulated white noise
White noise (NOISEX-92)	

Table 2.2: Noises used for training and testing SNR estimation algorithms

length of the speech were resampled to $f_s = 8\,\text{kHz}$ and mixed in SNRs calculated using the mean speech power derived from ITU-T P.56 [25] and the mean noise power. Noise levels between -5 and $36\,\text{dB}$ have been included in the experiments. Actual SNRs ranging from $-5\,\text{dB}$ to $35\,\text{dB}$ in $5\,\text{dB}$ increments were used for the test dataset, and with $-6\,\text{dB}$ to $36\,\text{dB}$ in $5\,\text{dB}$ increments for the training dataset. The A-weighted SNR was also calculated and is shown in the overall comparison. The coefficients from the polynomial fit were then evaluated using the output from each algorithm under test. Fifth order polynomials were found to provide a useful improvement over lower orders, and better performing than cubic or spline interpolation in the experiments, whilst higher orders than five did not increase accuracy.

Tests were performed covering the permutations of algorithm, noise and ground truth SNR for each dataset. A total of 1800 noisy speech samples were tested with each algorithm for each of the test and training datasets and for SNR and A-weighted SNR comprising 14.400 tests in total. Also computed were the 95% confidence intervals for the SNR estimation errors in dB bounded by the 2.5^{th} and 97.5^{th} centiles.

The estimated Real-Time Factor (RTF) defined as processing time divided by the duration of the speech signal for each algorithm was determined by measuring the elapsed processing time using the Matlab *cputime* function for each call on a 2.3 GHz Intel i5

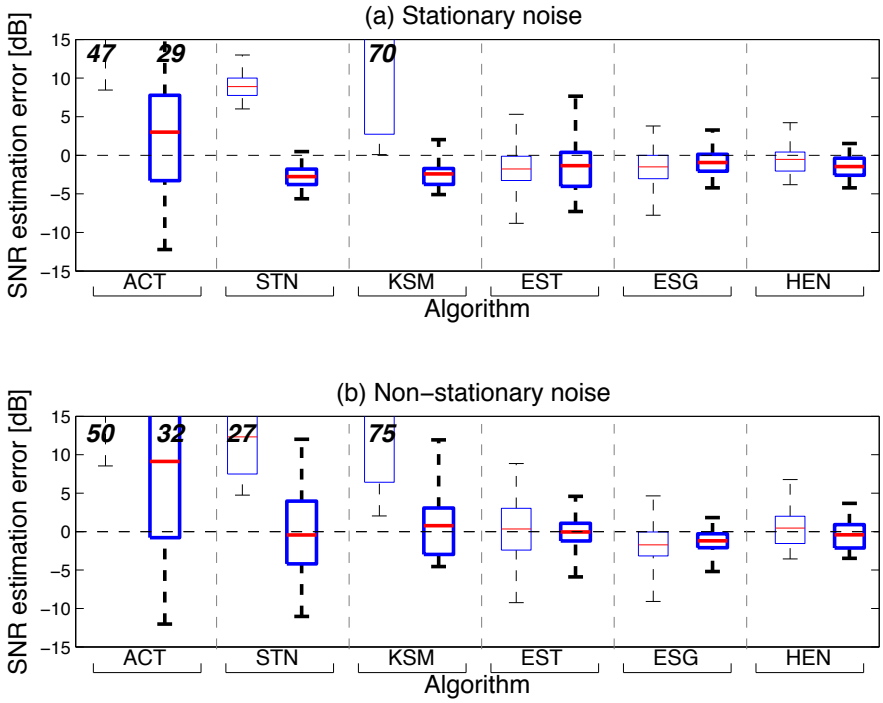

Figure 2.2: Box plots for SNR estimation error over the range -5 to $35\,\mathrm{dB}$ by algorithm before and after mapping. Results after mapping are shown in bold outline next to each unmapped result. Numbers in bold italics indicate extents of whiskers not shown

Core processor with 4 GB 1.333 GHz DDR3 memory, and calculating the mean time per algorithm divided by the mean speech file duration. This method of measurement provides only a rough insight into the absolute complexity of the algorithms, and does not give an indication of possible latency issues, but nevertheless provides useful insight into their relative complexity.

2.4 Results and discussion

2.4.1 Overall results by algorithm

The results for SNR estimation error by algorithm for SNR and A-weighted SNR before and after mapping are shown in Figs. 2.2 and 2.3 respectively. Results after mapping are plotted in bold. In the box plots [56], in each box the centre line shows the median, the top and bottom on the box show the 75^{th} and 25^{th} centiles respectively, and the whiskers extend to the 2.5^{th} centile and 97.5^{th} centiles between which 95% of the data lies. From

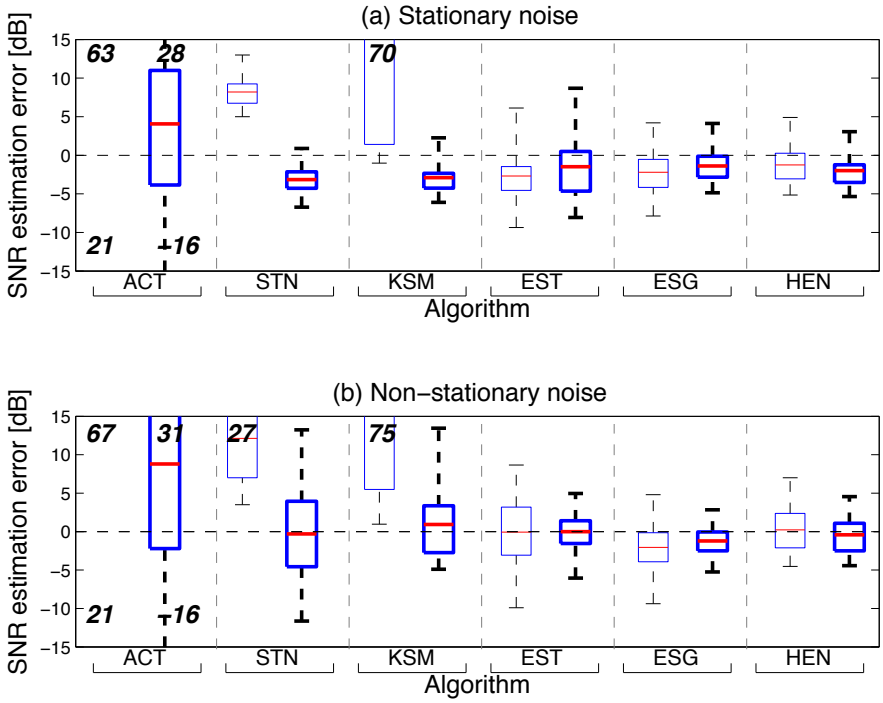

Figure 2.3: Box plots for SNR estimation error over the range -5 to $35\,\mathrm{dB}$ by algorithm with A-weighting before and after mapping. Results after mapping are shown in bold outline next to each unmapped result. Numbers in bold italics indicate extents of whiskers not shown.

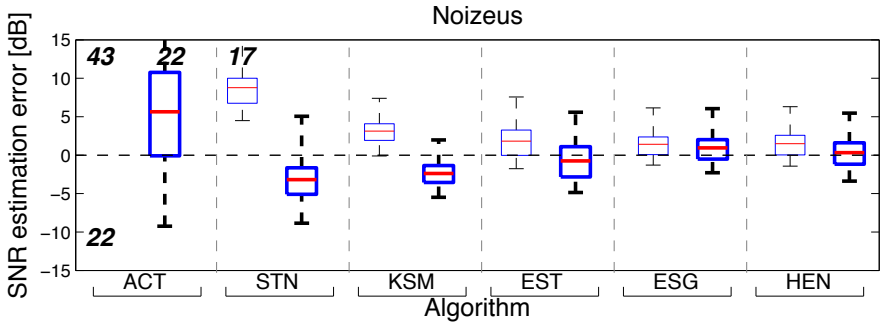

Figure 2.4: Box plots for SNR estimation error using the NOIZEUS corpus over the range 0 to $15\,\mathrm{dB}$ by algorithm before and after mapping. Results after mapping are shown in bold outline next to each unmapped result. Numbers in bold italics indicate extents of whiskers not shown

Fig. 2.2 it can be seen that the mapping technique substantially improves the accuracy and reduces the bias of all algorithms for both noise types. All algorithms produce large variances at low SNRs, and the mapping reduces this variance.

The ACT algorithm estimates have a large error as the algorithm estimates speech

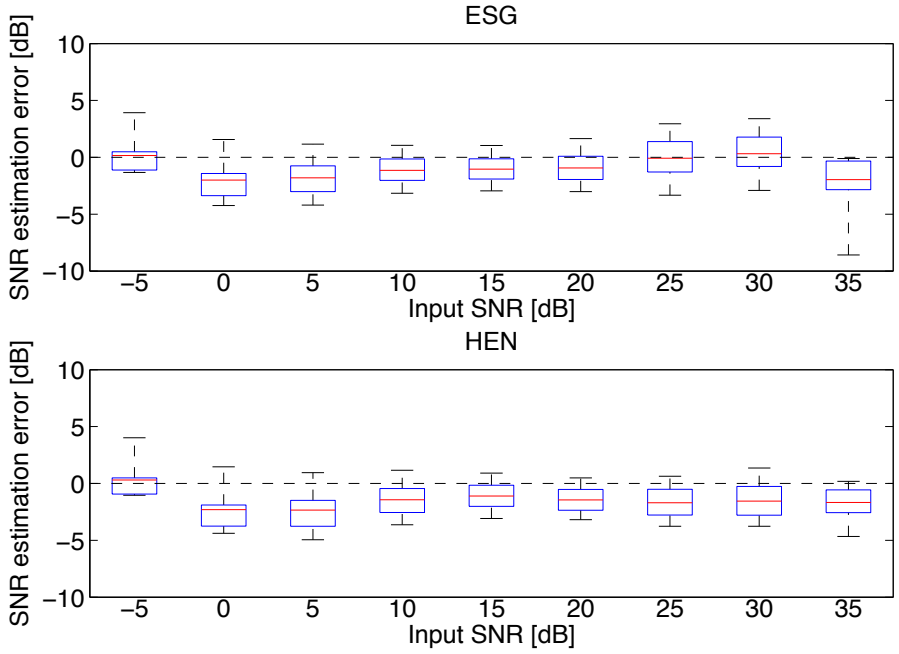

Figure 2.5: Box plots of estimation error by input SNR in stationary noise for the ESG and HEN algorithms after mapping

power including any noise present during active speech.

The STN algorithm performs well in stationary noise with 95% of estimation errors falling between 1.8 and −3.4 dB. With non-stationary noise however it performs poorly: the negative excursion at −5 dB not shown is approximately −25 dB.

The EST algorithm gives better performance on non-stationary sources whilst performing slightly less well on stationary data.

The HEN algorithm provides the best performance overall with a spread of data of 7.3 dB over 95% of results.

The ESG algorithm provides similar performance after mapping except at high SNRs but with much reduced computational complexity as will be discussed later in this section.

Figure 2.3 shows that the results for A-weighted SNR estimation are very similar to the unweighted results for all algorithms.

For further verification of the results, the algorithms were tested on the NOIZEUS [47] corpus using the polynomial mapping functions determined from the TIMIT speech corpus for each algorithm. The NOIZEUS corpus comprises speech in noise at SNRs of 0, 5, 10, and 15 dB. The noises used were: airport; babble; car; exhibition; restaurant; station; street;

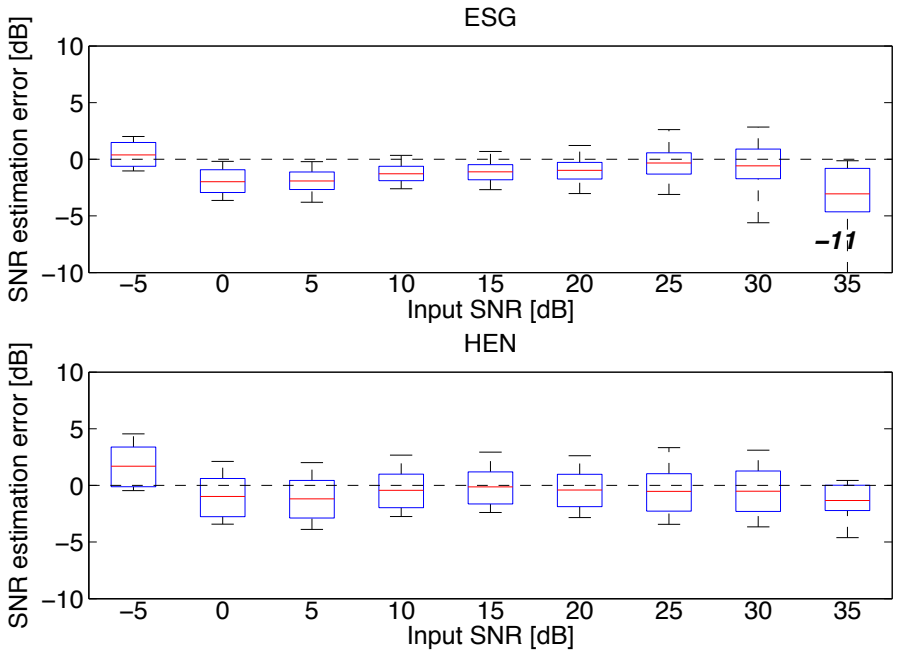

Figure 2.6: Box plots of estimation error by input SNR in non-stationary noise for the ESG and HEN algorithms after mapping. Numbers in bold italics indicate extents of whiskers not shown

and train; a mixture of both stationary and non-stationary noise. The 95% confidence intervals for the NOIZEUS tests are shown in Fig. 2.4. The SNR estimation performance ranking of the algorithms was similar to the TIMIT [19] stationary results, and as expected due to the mixture of stationary and non-stationary noise, algorithms which give large errors below 0 dB and above 15 dB in non-stationary noise show good results in this test.

2.4.2 Results by SNR

The SNR estimation errors for the two best performing algorithms, ESG, and HEN, after mapping, and calculated over the full range of test SNRs in stationary noise are shown in Figs. 2.5 and 2.6. The SNR after mapping is defined as $\hat{\xi}_{mapped}$. The characteristics and coefficients for $\hat{\xi}_{mapped}$ of the mapping functions based on training data for HEN and ESG are shown in Fig. 2.7. This illustrates the similar performance of HEN and ESG at SNRs of 15 dB and below, the very high accuracy of the HEN algorithm above 10 dB SNR, and the larger estimation errors at higher SNRs produced by the ESG algorithm. ESG is most accurate at 15 dB SNR, the optimal fixed prior $10\log_{10}(\xi_{\mathcal{H}_1})$ found in [30]. This prior is used to decide whether the noisy speech signal is more likely to be speech or noise, and the

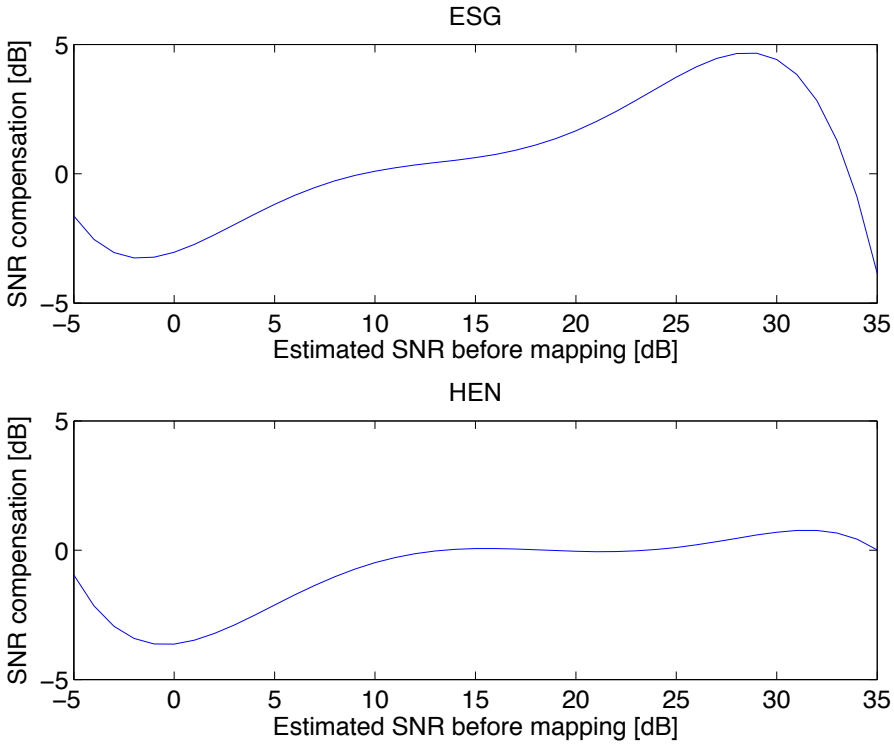

Figure 2.7: Mapping functions for the ESG and HEN algorithms

likelihood function is symmetrical either side of this value. HEN is least accurate at 0 dB, the point at which the bias function, B, in [29] increases.

The linear regression-based mapping functions here have been computed using SNRs between −5 and 35 dB, therefore the estimation performance outside of this range is uncertain.

2.4.3 Computational complexity

Table 2.3 shows the estimated RTFs for each of the algorithms using the method discussed in Sec. 2.2. The ESG algorithm has a similar RTF to STN but superior SNR estimation performance in non-stationary noise. The HEN algorithm has the highest RTF due to the safety net and bias compensation computations in the algorithm. In most circumstances therefore ESG may therefore be preferable to HEN except where good performance at high SNRs is required. All implementations were in Matlab except for STN which was a 64-bit compiled executable.

ACT	STN	KSM	EST	ESG	HEN
0.027	0.0045	0.0080	0.012	0.0053	0.35

Table 2.3: RTF by algorithm

2.5 Conclusion

A quantitative comparison of non-intrusive SNR estimation algorithms has been presented and the mapping function introduced. Tests were conducted over a wide range of input SNRs and noise types including both stationary and non-stationary noise. The experimental results show that for narrowband speech in noise the best performing algorithm is HEN [29], which has an ability to estimate noise to within approximately ± 3.5 dB within the range of input SNRs from -5 to 35 dB. The ESG algorithm however with only slightly reduced accuracy to HEN has significantly lower computational complexity. The use of a polynomial mapping function that can reduce both the bias and variance of the measurement error has been demonstrated, and that with the inclusion of this mapping function, 95% confidence intervals are reduced by around 5 dB for the ESG algorithm and 3 dB for the HEN algorithm. The results are also encouraging for the intended application in this thesis of enabling estimation of other acoustic parameters in noisy conditions.

Chapter 3

Non-intrusive Reverberation Time estimation from degraded speech

3.1 Introduction

A SIGNAL radiating from a source at a given position in a room will follow multiple paths to any observation point. These paths comprise the direct path which is in a straight line from the source to the observation point, as well as reflections from the walls, floor, ceiling and the surfaces of other objects in the room. The combination of the reflections forms the reverberation at a given observation point.

A sampled noisy reverberant speech signal, $y(n)$, captured by a microphone in a room is characterised by the Acoustic Impulse Response (AIR), $\mathbf{h}$, of the acoustic channel between the source and microphone. It is assumed that the AIR is linear and time invariant, and that the coefficients, $\mathbf{h}$, tend to zero as n increases thus only the first L_h samples are necessary to consider. Using a vector formulation, the noisy reverberant signal incorporating the convolution of the speech with the AIR can be written as

$$y(n) = \mathbf{h}^T \mathbf{x}(n) + v(n), \tag{3.1}$$

where

$$\mathbf{h} = [h_0, h_1, \ldots, h_{L_h-1}]^T \tag{3.2}$$

is the AIR of length L_h,

$$\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-L_h+1)]^T \tag{3.3}$$

is the speech signal vector, and $v(n)$ is the additive noise at the microphone at discrete time n.

Reverberation time is a fundamental parameter for quantifying room acoustics. A sound wave in a room will travel from the source and arrive at the various surfaces in the room before being reflected from those surfaces. The reflected sound waves will continue to travel around the room until all their energy has been dissipated. The absorption of each surface in the room, and the size of the room will influence the reverberant decay observed in the room.

Sabine observed that a sound source becomes inaudible when it has decayed by 60 dB after the source has been switched off [57]. Therefore, Reverberation Time (T_{60}) is defined as the time in seconds for a sound in the diffuse field to decay by 60 dB from the time the source has abruptly ceased, and can be characterized by the Sabine or Eyring equations [57, 58].

Sabine observed that the T_{60} was proportional to the volume of the room, V, in m^3 and also varied in inverse proportion to the equivalent absorption surface area in the room, A, in m^3, given by

$$A = \alpha_S \cdot \mathcal{S}, \tag{3.4}$$

where $\mathcal{S}$ is the absorbing surface area in m^2, and α_S is the Sabine absorption or attenuation coefficient. In Eyring's formulation, A is given by

$$A = \log(1-\alpha_E) \cdot \mathcal{S}, \tag{3.5}$$

where α_E is the Eyring absorption coefficient. Therefore, the T_{60} is given by

$$T_{60} = \frac{24 \cdot \log(10)}{c} \frac{V}{A}, \tag{3.6}$$

where c is the speed of sound in ms^{-1}.

The T_{60} is a function of room geometry and absorption and is nominally the same

throughout a given room in contrast to the AIR, which is dependent on the source-microphone configuration used for measurement. It can be expressed in FB or in frequency bands such as ISO-octave or ISO-one-third-octave [59].

The T_{60} estimate of a room is useful because it can provide an indication of the intelligibility and quality of speech observed in that room, and can be used to improve the performance of speech processing applications such as speech recognition [60] and speech dereverberation [17, 61, 62]. Standardized methods exist for estimating T_{60} from a measured AIR such as [63, 64]. More recently, the exponential sine sweep method [65, 66], has facilitated the measurement of AIRs by providing a means to provide robustness to noise. In many practical situations, however, measurement of the AIR is not possible, and the T_{60} must be estimated from reverberant speech non-intrusively, that is to say without prior knowledge of the AIR, the clean speech, or the SNR.

A further important acoustic characteristic of a room is the level of reverberation relative to the direct signal, which in combination with the T_{60} determines the overall envelope of the decay of the reverberation following the abrupt cessation of a signal. An objective measure of this ratio is the DRR which can be computed using the method of Mosayyebpour *et al.* [67] as

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=n_d-n_0}^{n_d+n_0} h_n^2}{\sum_{n=0}^{n=n_d-n_0} h_n^2 + \sum_{n=n_d+n_0}^{n=\infty} h_n^2}\right), \tag{3.7}$$

where the direct path signal arrives at sample time n_d, and $n_0 = 120$ samples, or 2.5 ms. The value of n_0 chosen must be greater than zero since it is not possible to precisely determine the direct path when estimating the DRR from measured AIRs. The value of 2.5 ms represents an additional path difference of 0.85 m, for $c = 340\,\mathrm{m\,s^{-1}}$, where c is the speed of sound, and therefore includes the direct path but none of the early reflections from the walls of the rooms. It is necessary to include the energy before the arrival of the direct path because of the influence of the sampling process.

An important diffference between the T_{60} and the DRR is that whilst T_{60} is nominally the same for an entire room, DRR as with the AIR is dependent on the locations of the source and the observation. This chapter is concerned with T_{60} estimation, but since T_{60}

and DRR are related parameters, DRR is introduced here since assumptions about DRR are necessary when considering T_{60}. Non-intrusive DRR estimation is the subject of further study in Chapter 4.

3.1.1 Methods for non-instrusive T_{60} estimation in the literature

A common T_{60} estimation method is to identify periods where there are speech end points, and estimate the slope of the decays followed by statistical analysis either in full-band or in frequency bands. Lebart *et al.* [61] as part of a de-reverberation algorithm proposed a method based on detection of the longest speech pauses and estimation of the slope of the logarithm of the smoothed energy envelope through linear regression. Although the test cases were limited, accurate results were obtained with an Root Mean Square (RMS) error of 24.7% in noise-free conditions. Baskind *et al.* [68] proposed a method based on automatic detection of decaying segments followed by Schroeder integration and linear regression. Ratnam *et al.* [69] modelled the decays using an exponential function with a time constant based on the the reverberant speech signal. A Maximum Likelihood (ML) fit is then made for each speech endpoint which is then used to compute the T_{60}. Ratnam *et al.* [70] subsequently improved on the performance of this method. Vieira *et al.* [71, 72] restricted the reverberation modelling process to the Free-Decay Regions (FDRs) where there is a consistent reduction in sound energy over consecutive frames. Vesa *et al.* [73] proposed a method requiring a binaural speech signal. Suitable sound segments for analysis are located using short-time energy and inter-channel coherence measures, which is then followed by the Schroeder integration method, line fitting and finally statistical analysis. The slope is estimated in the region that maximizes the correlation coefficient of the least squares method.

On the assumption that the ML approach works well on noise-free speech, Zhang *et al.* [74] proposed using Blind Source Separation (BSS) to estimate the T_{60} from noisy reverberant speech. The interfering sound sources are reduced in amplitude using first BSS and then an Adaptive Noise Canceller (ANC). An ML approach is then used to estimate the T_{60}. In a single test with a T_{60} of 270 ms an estimate of 300 ms is obtained at an SNR of 0 dB.

Löllmann *et al.* [75–77] proposed a series of ML-based methods which attempt to

solve the problem of additive noise. Prior to estimation, denoising is performed, using for example spectral subtraction, and then the speech signals are downsampled to reduce computational complexity. Pre-selection is used to detect the likely decays that are then used in an ML estimation process. A histogram is used in order to provide more robust estimates. Prego *et al.* [78] proposed an algorithm based on [71, 72] but using subband decomposition of speech signals to allow a much shorter reverberation speech signal to be used. Whilst performance exceeded that of the other methods in the comparison, the authors did not evaluate the performance of the algorithm when additive noise is present in the reverberant speech signal. Kendrick *et al.* [79] proposed a method based on [69] but where a model of non-diffuse sound decay was applied. This allows the decoupling of early and late reverberation decay regions and the removal of low dynamic range decay phases. Kendrick tested with SNRs down to 5 dB in an evaluation specific to classrooms and hospital wards where a very long recording can be made allowing an accurate estimate. It was also found that at least 40 hours of recording were needed in a hospital to produce and accurate Reverberation Time (T_{20}) estimate. This is not practical for a real-time or near real-time implementation. Jan *et al.* [80] proposed a method assuming a Laplace distribution for the sound decay rather than Gaussian, and an ML-based method for parameter estimation. The method was tested on AIRs without noise and achieved comparable to results to other algorithms not focused on the noisy case. Diether *et al.* [81] proposed an ML method aimed at estimating T_{60} up to 6 s in frequency bands and claims state-of-the-art performance for T_{60} up to 1.5 s.

Wu *et al.* [82] proposed a method based on pitch strength of voiced speech using a pitch tracking algorithm. Pitch strength is used to find the cleanest signals across all frequency channels from a gammatone filterbank. Correlograms are used on each of the 55 channels to find those least corrupted by noise. The authors found a monotonically increasing relationship to T_{60} between the detected pitch periods provided by the pitch tracker, and the time lags from the correlograms. The method yields an estimate of T_{60} up to 0.6 s. The estimator is however adversely affected by the gender of the talker, and a different relationship exists for female speakers than male due to the differing harmonic content.

Temporal dynamics information has been used in a number of approaches using modulation frequencies. Unoki *et al.* [83] proposed a method based on the Modulation

Transfer Function (MTF) concept using inverse-MTF filtering in the modulation frequency domain. The test database comprised speech from [84] uttered by a single female talker. Falk *et al.* [85] proposed a method using short-term and long term temporal dynamic information. Short-term temporal dynamic information obtained from differential (delta) cepstral coefficients and long term temporal dynamic information obtained from the modulation spectrum is used as input features for T_{60} estimation. The authors demonstrate performance exceeding their baseline system which was found to be more robust to background noise than [69,75]. Their proposal also estimates DRR. Falk *et al.* [86] proposed a further method based on the Speech-to-Reverberation Modulation Energy Ratio (SRMR) using the modulation spectrum. Using an auditory-inspired filterbank analysis of critical-band temporal envelopes of the speech signal, the method analyses the energy in eight modulation frequency bands. The SRMR is the ratio of the average energy in the low modulation frequencies (4 Hz to 18 Hz) to the high modulation frequencies (29 Hz to 128 Hz), and it is shown that the reciprocal of the SRMR is highly correlated with the T_{60}.

An alternative approach is to examine the decay rates of the entire speech signal rather than by examining the speech pauses alone. The Spectral Decay Distributions (SDD) method first proposed by Wen *et al.* [87] analyses the decay rates of the reverberant speech signal in time-frequency bins. The authors observed a strong correlation between the negative decay rates of the speech signal and the reverberation time, and by using a polynomial mapping function, the method provided a means of estimating the reverberation time. The method was however implementation dependent and needed to be trained on a suitable speech corpus. The estimation was also only valid in when the microphone was beyond the critical distance in the diffuse field where DRR < 0 dB. Lopez *et al.* [88] proposed the addition of pre-whitening filter to the SDD method to remove the speech components from the signal leaving only the Linear Prediction (LP) residual which is suggested contains most of the information regarding the AIR. The results show that the pre-whitening filter reduces the variance of SDD, however this method is tested on clean speech only and does not investigate the noisy case. Talmon *et al.* [89] proposed an SDD-based method using a data-driven approach to rank a series of training rooms via the eigenvalue decomposition of a specially tailored kernel, which was then extended to estimate the decay rate of a new room. The authors suggested that the technique orthogonalises the speech, noise and reverberation components and results in very accurate estimates in

high levels of noise. Dumortier presented an SDD-based method which uses a two-channel null-steered beamformer to remove the direct path components of the signal thus providing improved accuracy over [87], and without the requirement for the microphone to be beyond the critical distance [90].

A further approach was suggested by Lim *et al.* [91, 92] where techniques used in image processing to determine blurring are applied to the spectrogram of the speech signal. The amount of blurring in the image is shown to be related to the reverberation time.

Table 3.1 shows relevant studies in chronological order, and which data was used (where available) in the papers reviewed.

Study	Method	Speech source (language)	AIR source	Noise source
Cox [93]	Neural networks	Not available	Not available	N/A
Lebart [61]	ML	Not stated, French	Measured	N/A
Baskind [68]	ML	Monophonic	Not available	N/A
Ratnam [69]	ML	Connected Speech Test (CST) [94], International Phonetic Association (IPA) word utterances	Measured	N/A
Ratnam [70]	ML	IPA word utterances	Polack [95]	N/A
Vieira [71]	ML	Not available	Measured	N/A
Vieira [72]	ML	Not available	Measured	N/A
Vesa [73].	ML	Non-speech	Measured	N/A
Zhang [96]	BSS and ML	M.	Source-image method [97]	40s male utterance
Wu [82]	Pitch strength	Randomly selected from TIMIT [19], M,F, English	Source-image method [97]	N/A

Study cont.	Method cont.	Speech source (language) cont.	AIR source cont.	Noise source cont.
Wen [87]	SDD	Not stated	Measured, Polack [95], Source-image [97]	N/A
Unoki [83]	MTF	Advanced Telecommunications Research Institute International (ATR) Speech Database [84], Japanese, F.	Polack [95]	N/A
Löllmann [75]	ML	Not stated	Polack [95]	Babble from NOISEX-92 [20]
Löllmann [76]	ML	Not stated	Polack [95]	NOISEX-92 [20]
Falk [85]	Cepstral coefficients and modulation spectrum	M,F, English,French	Simulation of REal Acoustics (SIREAC) [98], ITU-T G.191 [99]	SIREAC [98]
Falk [86]	SRMR	Continuous Speech Recognition Wall Street Journal Phase 1 (CSR-WSJ) [100] M,F, English	Multidimensional Colouration Space (MCS) [101] Multichannel Acoustic Reverberation Database at York (MARDY) [102]	NOISEX-92 [20]

Study cont.	Method cont.	Speech source (language) cont.	AIR source cont.	Noise source cont.
Löllmann [77]	ML	Not stated	Jeub 2009 [103]	Not stated
Prego [78]	ML	Not stated	Source-image method [97], Measured, Jeub 2009 [103], [104], MARDY	N/A
Kendrick [79]	ML	Own recordings, M., English	Measured	2-5 additional speakers
Jan [80]	ML	TIMIT [19]	Jeub 2009 [103]	N/A
Lopez [88]	SDD with pre-whitening	CMU Arctic [105]	Fast Image Source Method (FISM) [106]	N/A
Talmon [89]	SDD with Intrinsic modelling	Not stated, M,F.	Source-image method [97]	White Gaussian Noise (WGN)
Dumortier [90]	SDD with null-steered beamformer	Grid [107]	Source-image method [97]	WGN
Lim [91]	Blur kernel	TIMIT, Gender not stated	Aachen [103]	WGN
Lim [92]	Blur kernel	Acoustic Characterization of Environments (ACE) [6], M,F	ACE [6]	ACE [6] (Ambient, fan, babble)

Study cont.	Method cont.	Speech source (language) cont.	AIR source cont.	Noise source cont.
Diether [81]	ML	Not stated, M,F	Open AIR Library [108]	N/A

Table 3.1: Speech, AIR, and noise corpora by algorithm

3.1.2 Estimating T_{60} in noise in the literature

Proposals prior to Löllmann *et al.* [75] did not tackle the noisy case where there is additive noise mixed with reverberant speech. Subsequently, Gaubitch *et al.* [109] conducted an evaluation of state-of-the-art methods [77, 86, 87] under noisy conditions and showed that these methods exhibit a large bias in the presence of high levels of noise. Gaubitch *et al.* [109] also found [85] to be robust to noise but with a large variance with respect to different speakers. A direct comparison of the algorithms [80, 88, 89] has yet to be conducted and there is an opportunity for such a study involving different speakers, SNRs, and noise types. Suitable databases to use for such a study include the TIMIT database for the speech since it contains a wide range of male and female speakers and both a test and training corpus and is widely used, and Database to Study the Effect of Additive Noise on Speech Recognition Systems (NOISEX-92) for the noise since it is widely used and contains a range of accepted noise types. Other non-stationary noises would be required to supplement NOISEX-92 since it contains largely stationary noises. Impulse responses from both simulated and real rooms should be used to demonstrate robustness. Impulse responses generated using Polack's method [95] and the Source-image method provide good simulations of small-room acoustics [97]. For the latter a flexible software tool can be used [110].

3.1.3 Aims and organisation of this chapter

Whilst the problem of T_{60} estimation in clean reverberant speech has been tackled in several different ways and with good accuracy, noise is increasingly present in all speech signals [22]. The problem of estimating T_{60} for noisy reverberant speech is still an open

research question, and addressing this problem is the aim of this chapter.

The key contributions of this chapter are therefore: to propose a T_{60} estimation method employing SDD which is substantially robust to additive noise thus avoiding problems of bias in existing estimators; to show how the cost of computing the SDD can be reduced; and to show a comparison of the improved method's results with previous research. The baseline method selected for comparison is the algorithm by Wen *et al.* [87] because it performed well in noise-free conditions, and showed little variance with different speakers in the detailed benchmarking [109].

The subject of a thorough evaluation of non-intrusive T_{60} estimation algorithms is tackled in Chapter 6 where a corpus of speech, noise and AIRs is constructed [6] and then used to conduct an evaluation.

The remainder of this chapter is organised as follows: In Section 3.2 the baseline T_{60} estimation method, SDD, is reviewed. In Section 3.3 the proposed approach to reducing the computational cost of the SDD calculation providing robustness to additive noise is discussed. In Section 3.4 the experimental approach to testing the proposed algorithm is discussed. The results are discussed in Section 3.5, and presented in Section 3.6 are the conclusions.

3.2 Review of SDD T_{60} estimation

3.2.1 Operating principle

Consider speech observed in an anechoic room. The amplitude envelope of the speech will have many different attack and decay phases. The rates of these attacks and decays will differ from word to word, and there will be a large variation in speech decays rates in each sentence of speech. Now consider the same speech observed in a reverberant room. The reverberation in the room will extend the observation of each phoneme such that each decay phase will be extended by the reverberant decay, changing the observed decay rate. Since all speech observed in the room will be affected by the reverberation, it will also reduce the variation of the speech decays rates. In a room with a very large amount of reverberation, the decay rates will be such that there is almost no variation in decay rates between one word and another. This variation in decay rates can therefore be used to

estimate the reverberant decay in a room, and is the principle upon which the SDD method is based.

3.2.2 Room decay model

Room reverberation consists of direct sound, early reflections and late reverberation. The fine structure of late reverberation is typically modelled statistically whilst the decaying envelope of the AIR can be modelled as a deterministic signal parameterized by some damping constant, δ [95, 111]. This model assumes that the sound field in the enclosure is diffuse and the source-microphone distance is greater than the critical distance [111]. Polack developed a time-domain model that describes the AIR as one realisation of a non-stationary stochastic process [95] described by

$$h(t) = b(t)e^{-\delta t} \quad \text{for} \quad t \geqslant 0, \tag{3.8}$$

where $b(t)$ is zero-mean stationary Gaussian noise, and damping constant, δ is related to the reverberation time, T_{60} by

$$T_{60} = 3 \ \log 10/\delta. \tag{3.9}$$

The room decay model can be defined using (3.8) as

$$E\{h^2(t)\} = \sigma_b^2 e^{-2\delta t} = \sigma_b^2 \mathrm{e}^{\lambda_h t}, \tag{3.10}$$

where σ_b^2 denotes the variance of $b(t)$, and the decay rate, $\lambda_h = -2\delta$. The decay rate, λ_h, in this context is the rate at which the signal level falls exponentially following an abrupt signal end-point. By extending (3.8), the frequency dependent room decay model can be expressed as

$$\bar{H}(t, f) = P(f)e^{\lambda_h(f)t} \quad \text{for} \quad t \geqslant 0 \tag{3.11}$$

where $\bar{H}(t, f)$ is the energy envelope of AIR at time t and frequency f, $\lambda_h(f)$ is the decay rate at frequency f, and $P(f)$ is the initial PSD of the reverberant speech signal at frequency f. Equation (3.11) can be linearized by taking the natural logarithm:

$$\log \bar{H}(t, f) = \log P(f) + \lambda_h(f)t \quad \text{for} \quad t \geqslant 0. \tag{3.12}$$

The decay rate $\lambda_h(f)$ can therefore be estimated by applying a linear fit to the natural logarithm of the time-frequency energy envelope of the reverberant speech signal.

3.2.3 The SDD Method

The SDD method can be used as in [87] to provide a means of estimating T_{60} by observing the energy envelope of a reverberant speech signal in the diffuse field. Frequency-dependent decay rates are estimated for each analysis time frame by applying a least-squares linear fit to the log-energy envelope of the reverberant speech signal in each frequency band in the Short Time Fourier Transform (STFT) domain. As shown in [87] the Negative-Side Variance (NSV), defined as the variance of only the negative gradients in the distribution of the decay rates, correlates well with the room decay rate, and by using a polynomial mapping function, the NSV can be used as an estimator for the T_{60}. The mapping function must be trained on a suitable clean speech database that has been convolved with AIRs of known T_{60}. The NSV denoted by $\sigma^2_{x^-}$ is defined as the variance of a symmetrical distribution, $f_x^-(\lambda)$, with the same negative-side distribution of the original distribution, $f_x(\lambda)$, as in

$$f_x^-(\lambda) = \begin{cases} f_x(\lambda) & \text{for} \quad \lambda \leqslant 0 \\ f_x(-\lambda) & \text{for} \quad \lambda > 0. \end{cases} \tag{3.13}$$

Let $\lambda(k,l)$ equal the estimated decay rate for frequency band k and time frame l in a signal containing K frequency bands and L frames, and

$$\lambda'(k,l) = \lambda(k,l) \quad \text{for} \quad \lambda(k,l) < 0 \tag{3.14}$$

where only negative λ are relevant to decays. In the baseline approach of [87] the NSV is calculated as

$$\sigma^2_{x^-} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} (\lambda'(k,l))^2. \tag{3.15}$$

If the point of observation is closer to the source than the critical distance, the energy in the direct sound is greater than that of the reverberant signal, and for a speech endpoint, there will be an initial energy drop in transition between the direct sound and the reflected decaying sound. This will cause negative gradient estimates to be steeper and

will lead to an underestimation of the T_{60}. This can be mitigated using a multi-channel approach and a null-steered beamformer to remove the direct path as in [90].

As was observed in [109], the SDD method is susceptible to noise because the noise affects the gradient estimate of the decay rate in each time-frequency bin. Higher levels of noise will cause the gradient to tend to 0 biasing the estimate. Figures 3.1 to 3.5 show the gradients of the TIMIT speech signal TEST/DR1/FAKS0/SI1573 in time-frequency bins. Colours are used to represent the magnitude of the gradient of the signal in each time-frequency bin with brown representing a steep signal onset and dark blue representing a signal decay. In this representation, the largest magnitude gradients are likely to be speech onset and decays. Comparing Figs. 3.1 to 3.3, it can be seen that as the T_{60} decreases, the number of large magnitude gradients increases. This is because the reverberation will tend to fill the gaps between speech peaks and reduce the gradients within a given time-frequency bin. Comparing Figs. 3.3 to 3.5, it can be seen that as the SNR decreases

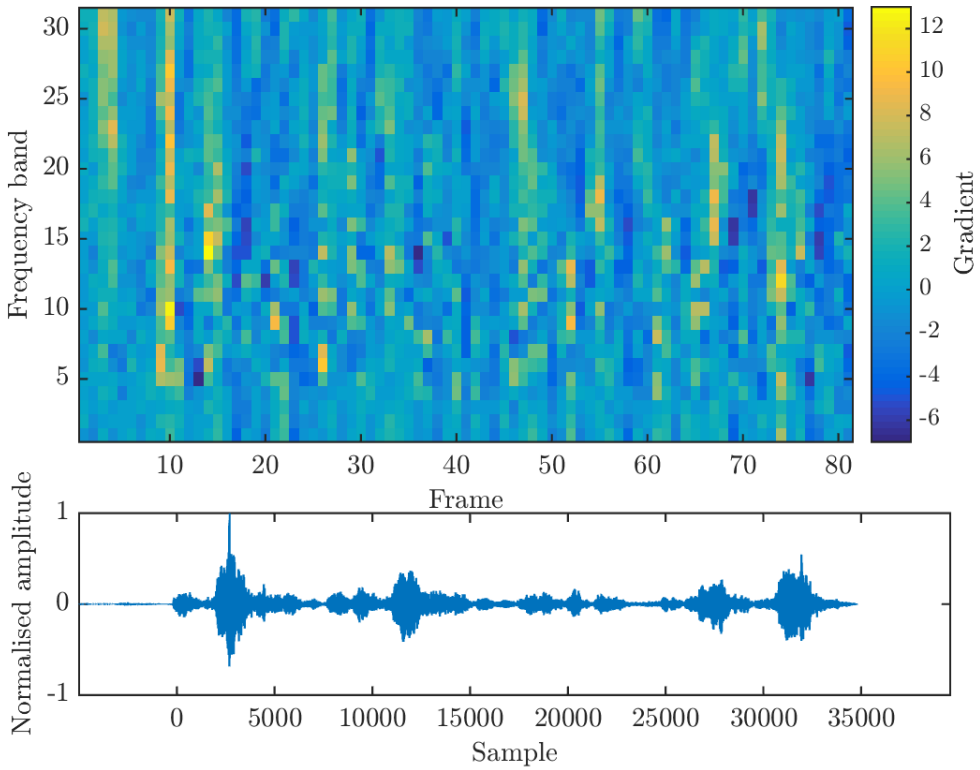

Figure 3.1: Time-frequency bin gradients for T_{60} = 600 ms, SNR = 35 dB

and the amount of additive noise increases, the number of large magnitude gradients decreases. Noise however decreases the number of large magnitude gradients much greater

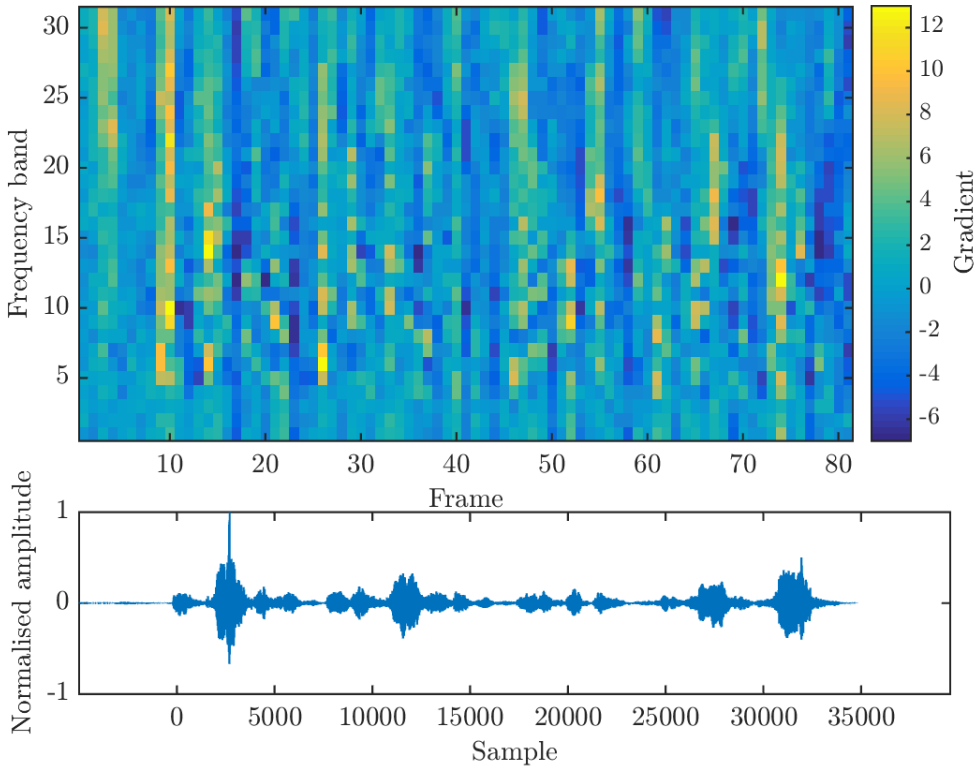

Figure 3.2: Time-frequency bin gradients for $T_{60} = 400\,\text{ms}$, SNR $= 35\,\text{dB}$

than T_{60} as shown by comparing Figs. 3.1 and 3.5. The consequences are that the SDD method grossly overestimates the reverberation time at low SNRs.

The SDD method was also shown in [109] to have a high computational cost relative to other estimators such as [77, 85]. This is because a least squares regression is required for each time-frequency bin and the large number of frequency bands used.

3.3 Proposed method

3.3.1 The SDD method with reduced computational cost

The proposed method follows a perceptually-motivated frequency analysis and employs a filter bank with uniformly spaced filters on the Mel frequency scale [112]. The number of Mel-spaced bands is much less than the number of STFT bands so that the computational complexity of the least squares fitting procedure in the proposed algorithm is correspondingly reduced compared to the baseline SDD. Additionally, since the Mel-spaced bands are formed

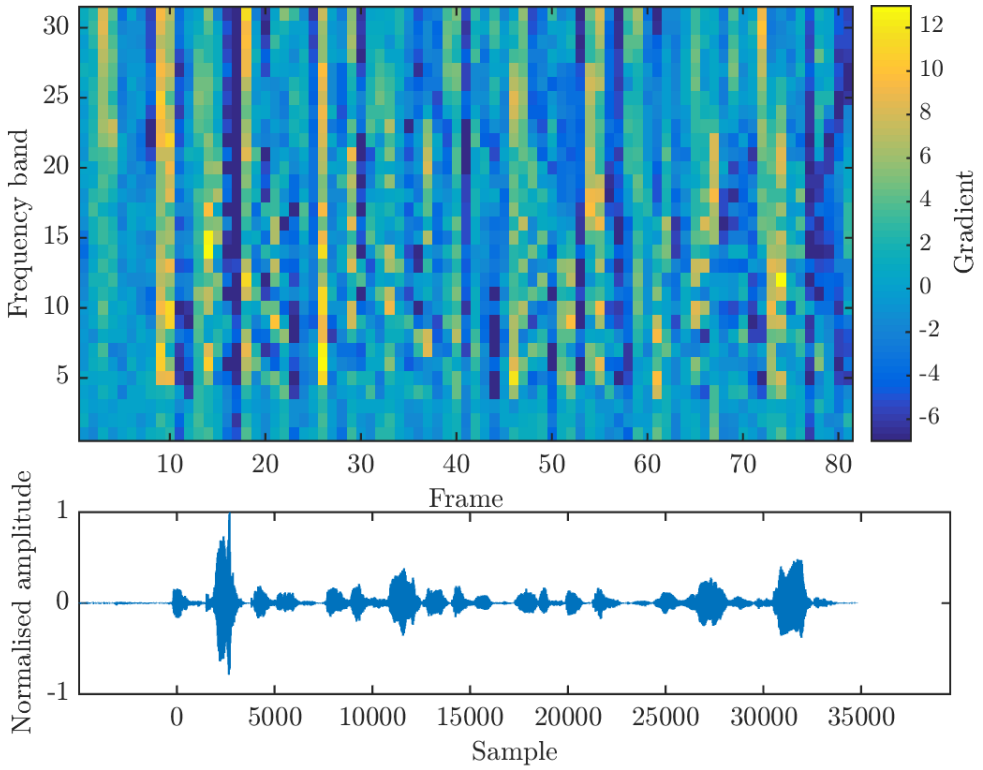

Figure 3.3: Time-frequency bin gradients for T_{60} = 200 ms, SNR = 35 dB

by weighted averaging of STFT bins, this also gives some reduced sensitivity to noise. It can be seen from the Bienaymé formula [113] for uncorrelated random variables

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i), \tag{3.16}$$

for a population with a small variance (i.e. the noise) and a population with a large variance (i.e. the speech) that in the limit where the ratio of the variances approaches infinity, the variance can be assumed to be the variance of the speech. This proposed algorithm is referred to as SDD with Mel-spaced frequency bands (SDDMSB).

3.3.2 SDD T_{60} estimation robust to noise

To reduce further the effect of noise on the T_{60} estimation variance, two additional concepts will be invoked in the algorithm for computing a representative decay rate for the signal given $\lambda(k, l)$, the raw decay rates estimated at each frequency band in each time frame:

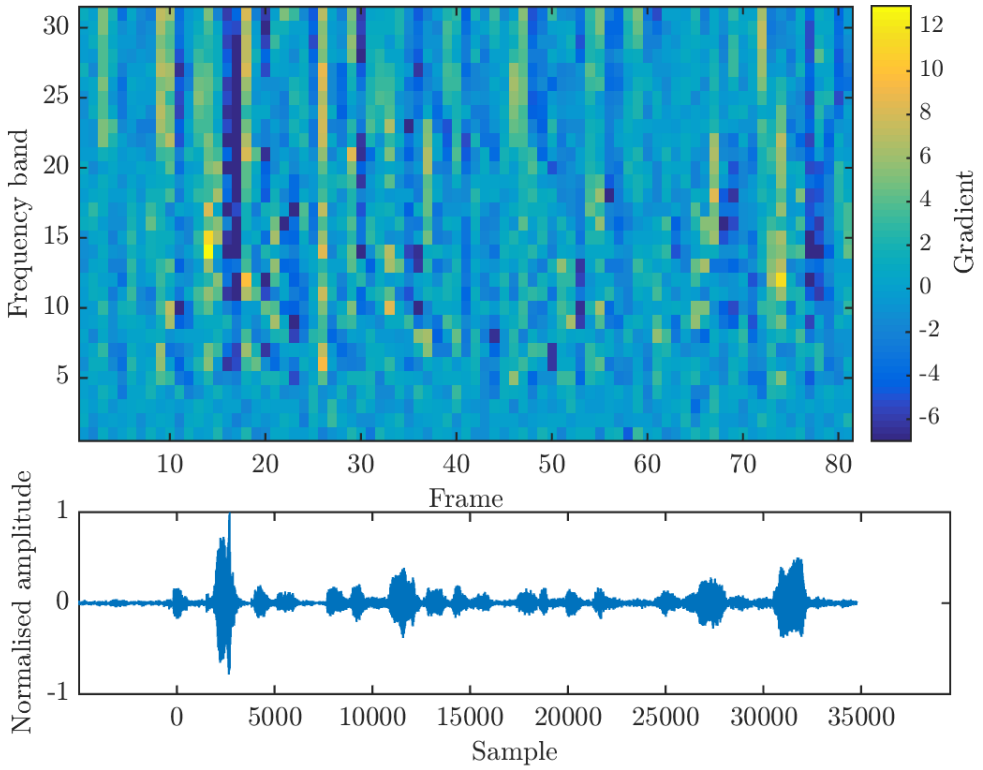

Figure 3.4: Time-frequency bin gradients for $T_{60} = 200\,\text{ms}$, SNR $= 15\,\text{dB}$

1. Selection of time-frequency bins where there is more likelihood of speech being present based on their decay rates for a given SNR

2. Switching of the selection method depending on the *a posteriori* SNR estimate

To illustrate this approach, consider time-frequency bins containing frames of the log magnitude speech spectrum averaged over Q Mel-spaced frequency bands near the fundamental frequency of the speech, typically between 85 and 255 Hz [114, 115]. The most negative gradients of these frames determined using least squares fitting for (3.12) will follow the decay of the speech with the reverberation tail. Now consider similar time-frequency bins containing noise that is uncorrelated with the speech and independent. The gradients of the frames containing noise will tend to zero. In the case of noisy reverberant speech, SDD estimation can be made more robust to the noise by basing it on the first of the above two cases, i.e. by selecting the highest negative decay rate gradients from time-frequency bins from (3.16).

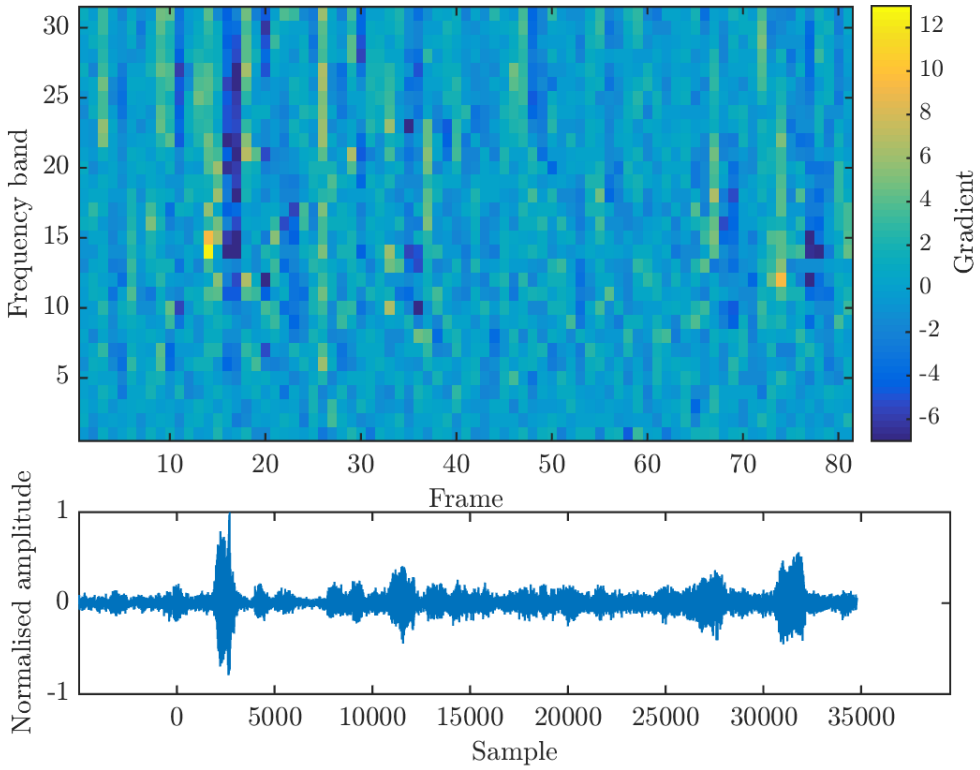

Figure 3.5: Time-frequency bin gradients for T_{60} = 200 ms, SNR = 5 dB

Next, mode-switching of the T_{60} estimation algorithm for operation in noise is employed, with switching controlled according to the input SNR. The choice of SNR threshold values in the following definitions of the modes will be explained below in Section 3.4.

- Mode A: for input SNRs better than 30 dB where the energy decay values across all frequency bands are used to determine the NSV and hence the T_{60} as in SDD.
- Mode B: for input SNRs between 15 and 20 dB a selection is made of energy decay values representative of clean speech is obtained by selecting the most negative gradient within each time frame across all frequency bands to estimate the NSV.
- Mode C: for input SNRs below 10 dB where noise power may exceed the speech power in a given time-frequency bin and produce large erroneous decays, the frames most likely to contain speech are identified under these conditions and it is assumed that the fundamental frequency of the speaker will tend to occupy one frequency band for most of the speech signal. To obtain the NSV in mode C the variance of each

entire frequency band for all frames is therefore computed, and the largest of these variances is selected.

The NSVs for modes A, B and C respectively are defined as

$$\begin{aligned} \sigma_A^2(\lambda) &= \sigma_{x^-}^2 \\ \sigma_B^2(\lambda) &= \tfrac{1}{L}\sum_{l=1}^{L}\left(\min_q(\lambda(q,l))\right)^2 \\ \sigma_C^2(\lambda) &= \max_q\left(\tfrac{1}{L}\sum_{l=1}^{L}(\lambda'(q,l))^2\right), \end{aligned} \tag{3.17}$$

where q is the index of the q$^{\text{th}}$ Mel frequency band. Averaging is employed in order to provide a smooth transition between modes, and also because the optimal SNRs for each mode are not contiguous. The estimated T_{60} is therefore given by

$$\widehat{T_{60}} = -3\log 10 \begin{cases} \frac{1}{m(\sigma_A^2(\lambda))} & \text{for} \quad \xi > 30 \\ \frac{2}{m(\sigma_A^2(\lambda))+m(\sigma_B^2(\lambda))} & \text{for} \quad 20 \leqslant \xi < 30 \\ \frac{1}{m(\sigma_B^2(\lambda))} & \text{for} \quad 15 \leqslant \xi < 20 \\ \frac{2}{m(\sigma_B^2(\lambda))+m(\sigma_C^2(\lambda))} & \text{for} \quad 10 \leqslant \xi < 15 \\ \frac{1}{m(\sigma_C^2(\lambda))} & \text{for} \quad \xi < 10 \end{cases} \tag{3.18}$$

where $m(\cdot)$ is a mapping function between the NSV and δ (as defined in (3.9)), and ξ is the SNR of the reverberant speech in decibels, which in practice is obtained from a noise estimator such as ESG discussed in Chapter 2. In tests, mode C was found experimentally to be most effective when used with at least 30 frequency bands in the Mel-frequency filter bank. Estimation accuracy of the algorithm was found to converge within three TIMIT utterances (approximately 8 s of speech) when trained on a single utterance. Therefore to be able to process realistic length audio streams, signals were split into 8 s blocks with T_{60} estimates found from the average across all blocks. Note that the consideration of bias within this estimator is beyond the scope of this chapter. This algorithm will be referred to as SDD with Mel-spaced frequency bands and selective averaging (SDDSA).

3.4 Performance evaluation

The proposed algorithm was trained and tested using simulated test signals at $f_s = 8\,\text{kHz}$. Speech signals were randomly selected from the training and test partitions of TIMIT [19] to produce exclusive training and test datasets. Speech signals were convolved with simulated AIRs, generated separately for training and testing using the source-image method [97, 110] for a room with dimensions $5 \times 4 \times 6\,\text{m}$, a source-microphone distance of 2 m, and T_{60} values from 200 to 950 ms in 150 ms intervals. The source position was $(2.0, 1.5, 2.0)$ and the receiver position was $(2.0, 3.5, 2.0)$. Training signals comprised single utterances from each of four different male and four different female speakers, whilst test signals comprised six utterances concatenated from each of four different male and four different female speakers to provide realistic tests on long sentences, and to avoid any per-speaker bias. The virtual source-microphone distance was always greater than the critical distance.

As mentioned in Section 3.2.2, the method assumes that the microphone is in the diffuse field. In practice this is achieved when the level of reverberation is greater than the level of the direct signal, that is to say that the DRR $< 0\,\text{dB}$. Table 3.2 shows the estimated DRRs of each AIR, $\mathbf{h}$, computed using (3.7). Since these DRRs are all $< 0\,\text{dB}$, the

T_{60} (ms)	200	350	500	650	800	950
DRR (dB)	−0.39	−7.0	−10	−12	−14	−15

Table 3.2: Estimated DRR of AIRs used in experiments

performance evaluation was conducted in the diffuse field.

Babble, White, Factory1 and Volvo noise signals, $v(n)$, from NOISEX-92 [20] were added to the reverberant speech test signals to simulate realistic noisy conditions. To obtain a noisy reverberant speech signal at the test SNR, ξ, each noisy reverberant speech test signal, $y(n)$, was constructed by convolving a clean speech signal, $x(n)$ with the respective AIR, $\mathbf{h}$, as in (3.1) to (3.3), and then mixed with the appropriate noise as

$$y(n) \quad = \quad \mathbf{h}^T \mathbf{x}(n) + \frac{v(n)\sqrt{\xi_s}}{\sqrt{\xi}} \tag{3.19}$$

where ξ_s is power ratio of the unprocessed speech and noise signals given by

$$\xi_s = \frac{A_x}{E\{|v(n)|^2\}}, \tag{3.20}$$

where A_x is the active speech level estimated using ITU-T P.56 Method B [25] of the entire reverberant speech utterance without noise, given by $\mathbf{h}_m^T\mathbf{x}(n)$, and $E\{\cdot\}$ is the expectation operator.

A one-time offline training procedure was used to determine the mapping function, $m(\cdot)$, for the relationship between the NSV and δ, as defined in (3.9), derived using a fourth order polynomial fit using the training dataset and the oracle T_{60} without noise. SDDSA was tested with the oracle SNR as well as the SNR from two state-of-the-art noise estimators: an implementation [37] of Gerkmann [30]; and Hendriks [29], henceforth referred to as SDDSA-G and SDDSA-H respectively. In addition, for each estimator the mode switching thresholds in (3.18) were optimized to minimize overall T_{60} RMS error in white noise. The risk of training the thresholds too specifically to the training data used was mitigated by reviewing the results across all noise types.

To evaluate SDDSA the SDD, SDDMSB, SDDSA-G, and SDDSA-H methods were compared in terms of RMS T_{60} estimation error as a function of either the oracle T_{60} or SNR. In addition, the estimated RTF for each algorithm was determined by measuring the elapsed processing time using the Matlab *cputime.m* function for each call and calculating the mean time per algorithm divided by the mean speech file duration. Tests were performed in Matlab on a 2.3 GHz Intel i5 Core processor with 4 GB 1.333 GHz Double Data Rate Type Three (DDR3) Synchronous Dynamic Random Access Memory (SDRAM).

3.5 Results and discussion

The performance of the T_{60} estimation algorithms were analysed using simulated data. First, SDDSA and SDDMSB with SDD were compared. Figures 3.6 and 3.7 show the RMS T_{60} estimation errors for the each method in Babble noise by SNR and T_{60} respectively. RMS T_{60} estimation error for SDD was typically over 200 ms rising to over 250 ms above 25 dB SNR whereas with SDDSA the RMS error was reduced to within 120 ms at 5 dB SNR and above. The justification and advantages of the switching scheme (3.18) can be seen in

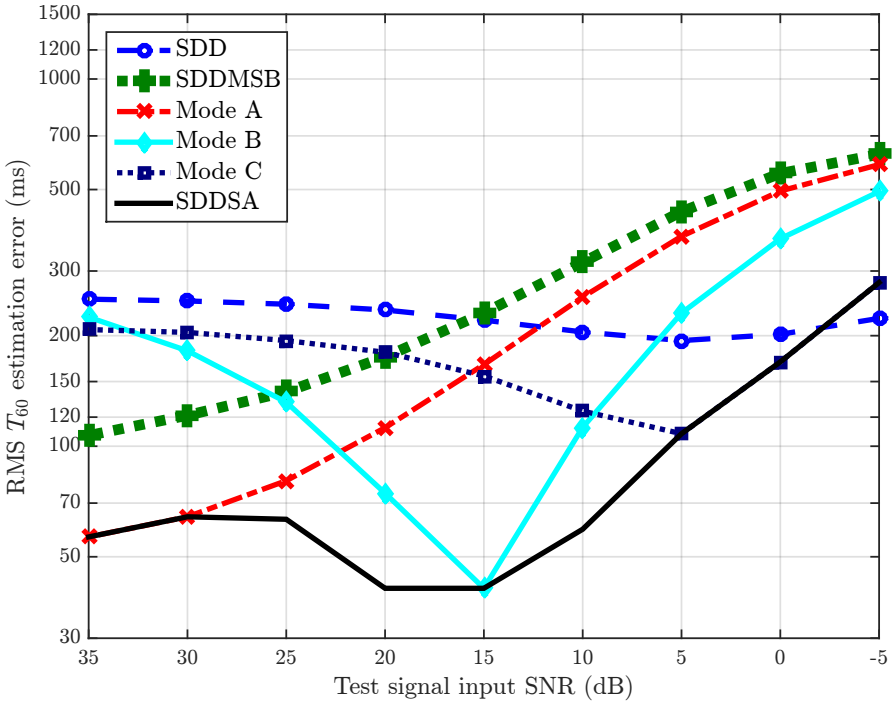

Figure 3.6: T_{60} **RMS estimation error by SNR and computation method in Babble noise using TIMIT speech in** T_{60}**s from** 200 **to** 950 ms **in** 150 ms **intervals**

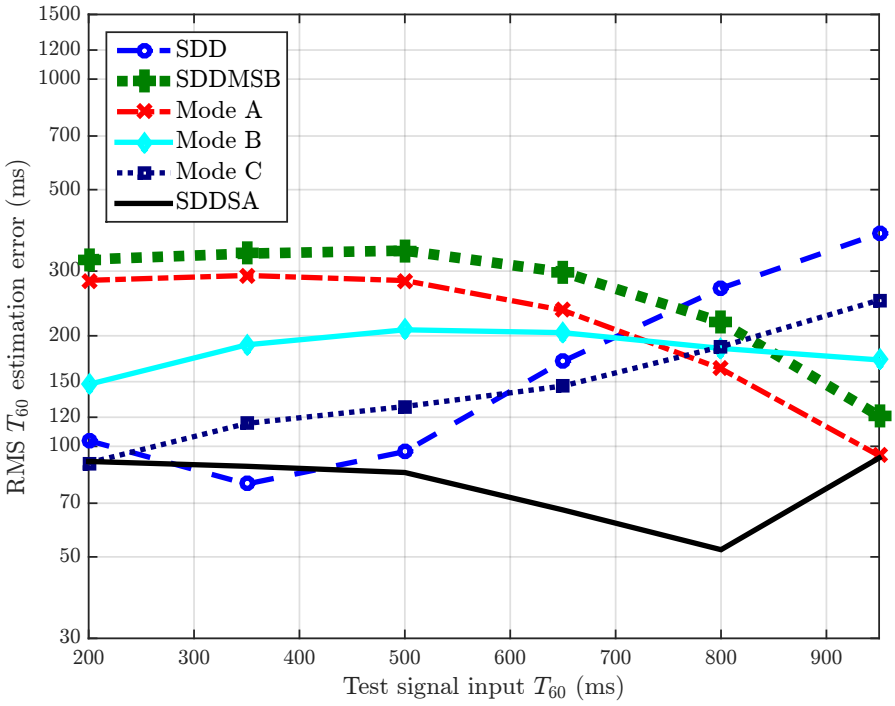

Figure 3.7: T_{60} **RMS estimation error by** T_{60} **and computation method in babble noise using TIMIT speech in SNRs from** 0 **to** 35 dB **in** 5 dB **intervals**

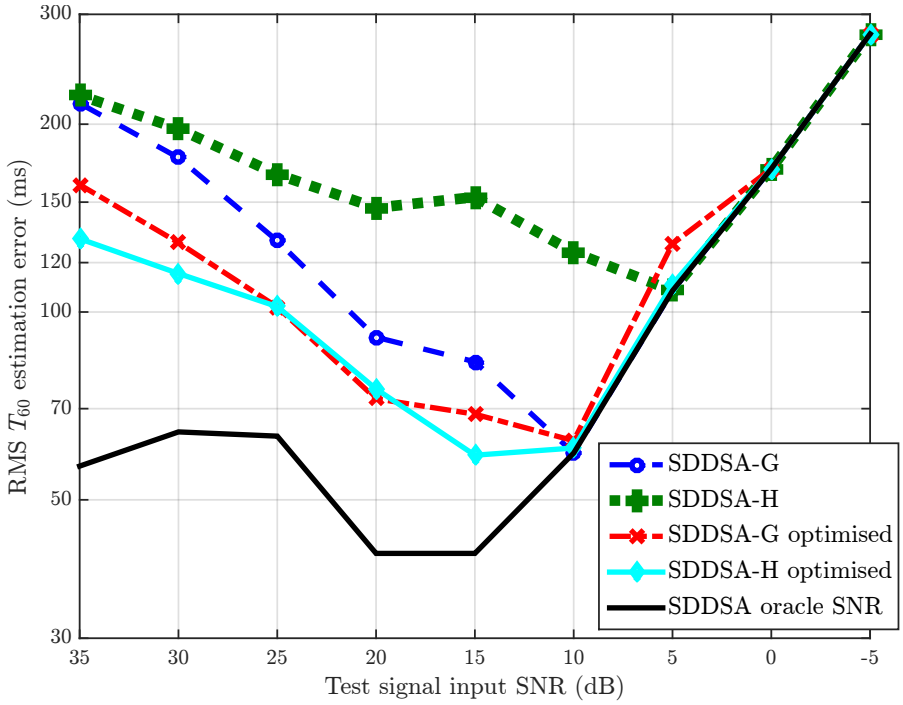

Figure 3.8: T_{60} **RMS estimation error by SNR and computation method in Babble noise using TIMIT speech in** T_{60}**s from** 200 **to** 950 ms **in** 150 ms **intervals**

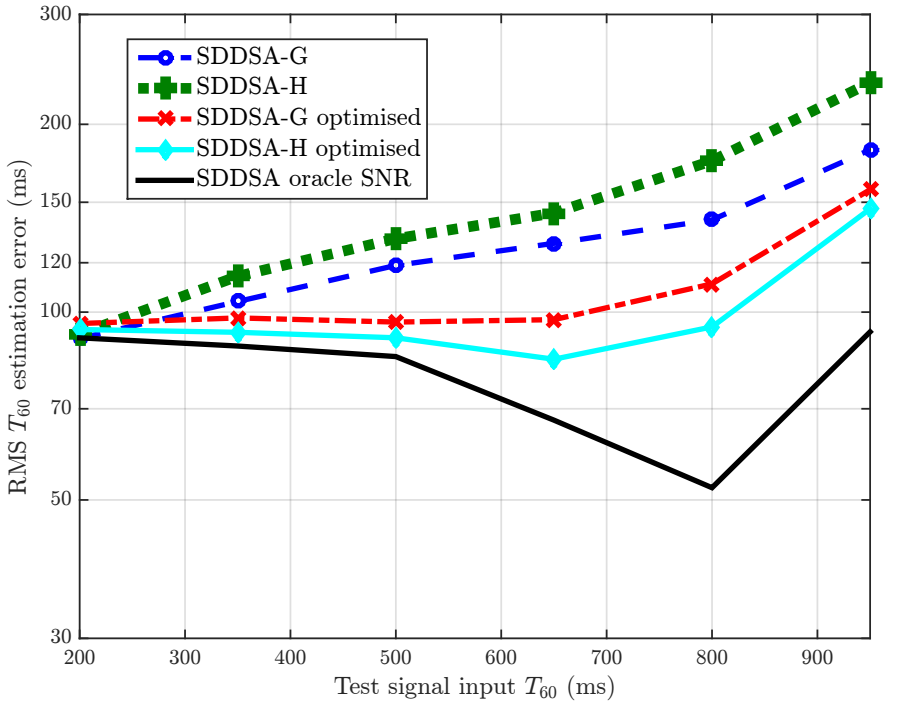

Figure 3.9: T_{60} **RMS estimation error by SNR and computation method in Babble noise using TIMIT speech in SNRs from** 0 **to** 35 dB **in** 5 dB **intervals**

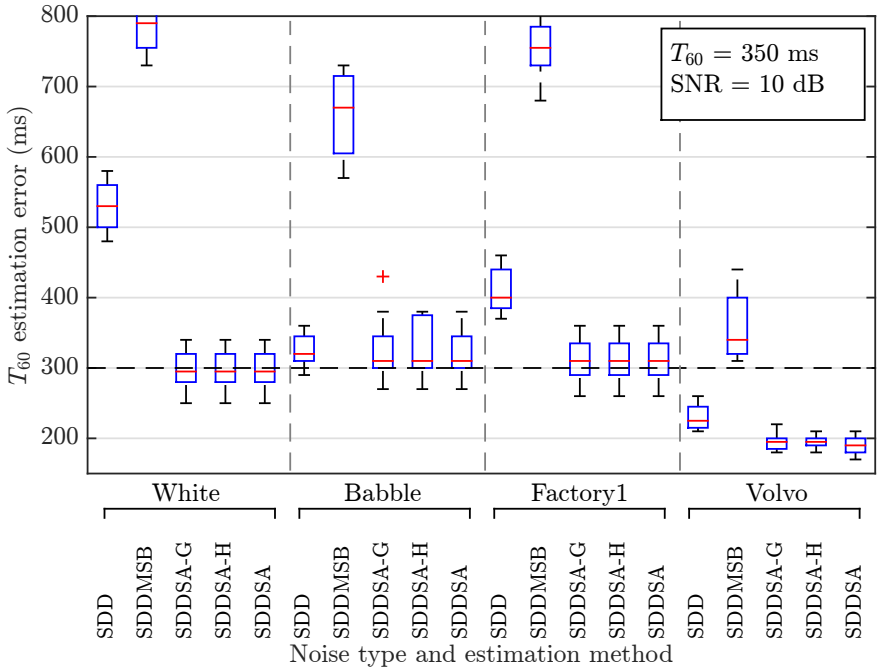

Figure 3.10: T_{60} estimation error by noise type and computation method at a specific operating point (T_{60} = 350 ms and SNR = 10 dB)

Fig. 3.6 which shows the T_{60} RMS estimation error for each mode as a function of SNR and that the lowest errors occur at different SNRs for each mode. Figures 3.8 and 3.9 show the estimation error of the SDDSA, SDDSA-G and SDDSA-H, the latter two both before and after threshold optimization. The case where the SNR is known *a priori* gives the lowest RMS T_{60} estimation error overall.

Using noise estimators with optimized thresholds brings the RMS T_{60} estimation errors to within approximately 50 ms of the case where the SNR is known *a priori*, with noise estimation errors having the greatest impact on the T_{60} estimate at SNRs of 20 dB and above. Noise estimation error is typically larger when the SNR is very large or very small hence the wide variation in results at 35 dB SNR shown in Fig. 3.8. Figure 3.10 shows the variation in estimation error at an example operating point with a T_{60} of 350 ms and an SNR of 10 dB in a range of noise types illustrating the effect of different speakers on the algorithm. On each box, the central mark is the median, the edges of the box are the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Errors greater than 800 ms are not shown since this level of error would not result in a reliable estimator.

Table 3.3 shows the RTF for each T_{60} estimation method. Whilst the RTF was larger than with the SDDMSB method, a fourfold ($4\times$) improvement over the SDD method is obtained. The higher RTF of SDDSA when compared to SDDMSB is due to the larger number of frequency bands needed to be able to differentiate the speech from the noise at low SNRs.

SDD	SDDMSB	SDDSA	SDDSA-G	SDDSA-H
2.97	0.26	0.67	0.68	1.51

Table 3.3: Comparison of RTF for T_{60} estimation methods

3.6 Conclusion

Non-intrusive estimation of reverberation time T_{60} from reverberant speech has been an important research topic for several years. It was shown in [109] that the method of Wen *et al.* [87] and two other state-of-the-art non-intrusive T_{60} estimation algorithms [77, 86] are strongly biased in the presence of additive noise. A novel SDD-based T_{60} estimation algorithm[1] has been presented that provides increased accuracy, reduced computational cost, and substantial robustness to additive noise with RMS T_{60} estimation errors of 120 ms or better for SNRs of 5 dB and above over a wide range of realistic noises degrading to around 250 ms at 0 dB. It has been shown that recent noise estimators such as Gerkman *et al.* [30] can be used instead of the oracle SNR without significantly degrading the T_{60} estimation accuracy. The algorithm is also the subject of a publication [1].

[1] Code available on http://www.commsp.ee.ic.ac.uk/~sap/projects/blind-estimation-of-acoustic-parameters-from-speech/blind-t60-estimator/ password:bBs96K

Chapter 4

Non-intrusive Direct-to-Reverberant Ratio estimation from degraded speech

4.1 Introduction

THE Direct-to-Reverberant Ratio (DRR), the level of an audio signal relative to the reverberation, introduced in Chapter 3, like the T_{60}, is another important parameter for quantifying room acoustics. As with the T_{60}, speech enhancement algorithm performance can typically be improved if the DRR is known [17], and together with T_{60} describes the envelope of the reverberation following the endpoint of an audio signal. An existing objective measure of the DRR is that of Mosayyebpour *et al.* [67] discussed in Chapter 3, Section 3.1. This method relies on information in the AIR regarding the onset and decay characteristics of the reverberation. However, when the AIR is not available, which is typically the case in many practical applications, the DRR must be estimated non-intrusively from the speech signal. The DRR is typically frequency-dependent [111], and an estimator that can also provide frequency-dependent DRR estimates could be used to provide an efficient dereverberation scheme for example by focusing the dereverberation on those frequencies where the DRR is low and thus problematical for speech quality and intelligibility.

4.1.1 Effect of DRR on a speech signal

Existing methods for non-intrusively estimating T_{60} from speech - a related problem to non-intrusive DRR estimation - rely on extracting decay information from the signal in frequency bands from speech endpoints as was discussed in Chapter 3. Whilst T_{60} is nominally the same for an entire room, DRR is dependent on the source-microphone configuration. In a single-channel signal, when the DRR is positive such as when the microphone is in the near field and the distance from the microphone to the source is less than the critical distance, there is a drop in the total signal level after an abrupt signal endpoint. In the far field however, when the DRR is negative, the amplitude of the of total signal (speech mixed with reverberation) will be greater than the speech signal. This is illustrated in Figs. 4.1 to 4.3 where four different T_{60} and DRR scenarios are considered using the word "sailboat" from the TIMIT training dataset utterance "A sailboat may have a bone in her teeth one minute and lie becalmed the next" .

Figure 4.1 (a) shows the word "sail" without reverberation. Fig. 4.1 (b) shows the same utterance where the T_{60} is low and the DRR is high, an acoustic environment typically found in small rooms where the talker is close to the microphone. The tail of the reverberant speech signal is very similar to the clean speech. In Fig. 4.1 (c), the T_{60} is high and the DRR is high, an acoustic environment typically found in large sparsely furnished rooms where the talker is close to the microphone. Both the tail and the body of the reverberant speech signal differs from the clean speech.

Figure 4.2 (a) again shows the word "sail" without reverberation for reference. In Fig. 4.2 (b) where the T_{60} is low and the DRR is low, an acoustic environment which might be found in small rooms where the talker is far from the microphone, there are significant differences in both the tail and the body of the reverberant speech compared to the clean speech. In Fig. 4.2 (c) where the T_{60} is high and the DRR is low, an acoustic environment typically found in large sparsely furnished rooms where the talker is far from the microphone, it is difficult to identify the onset or endpoint of the reverberant speech. The body is significantly shifted in time and the amplitude changed.

Figure 4.3 shows both the clean speech without reverberation (a) and the reverberant speech of the next frame in the sequence after Fig. 4.2. The endpoint of the word "sail" is completely obscured by the reverberation, and the word "boat" is combined with a large

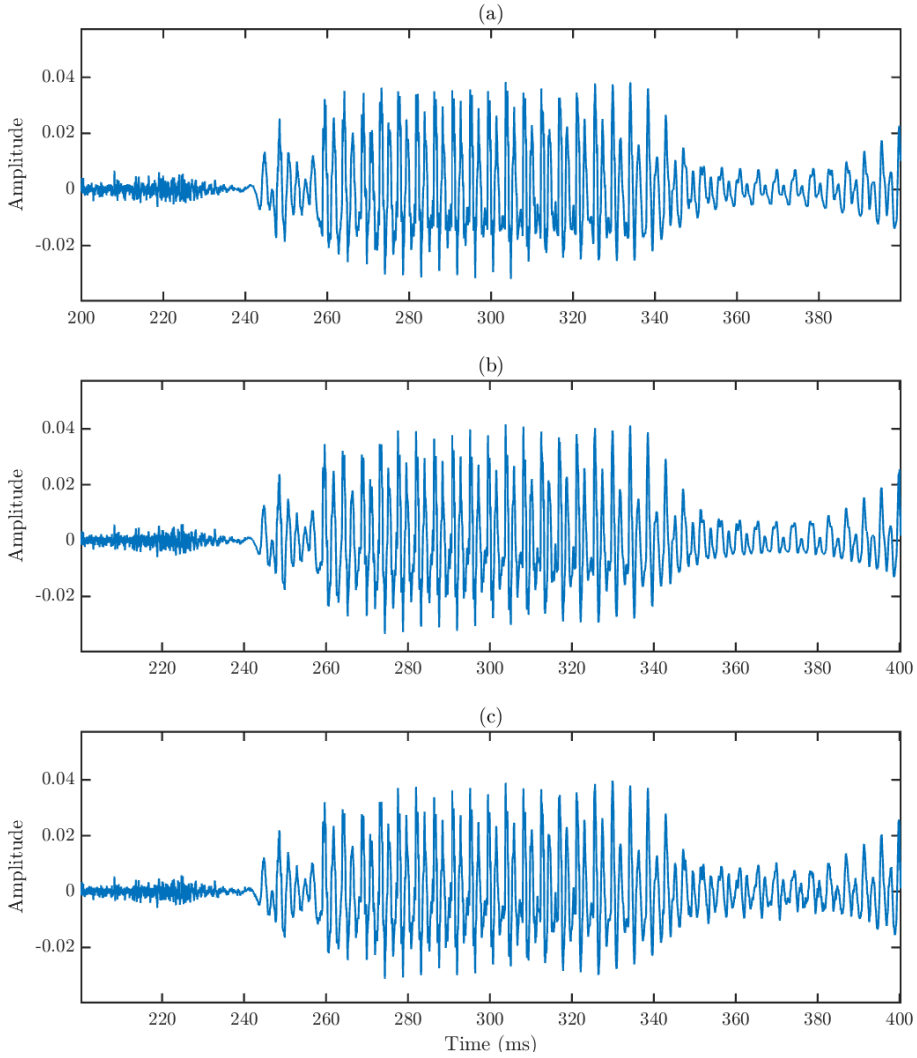

Figure 4.1: TIMIT training dataset utterance "A sailboat may have a bone in her teeth one minute and lie becalmed the next" in low and high T_{60} with high DRR scenarios. (a) is without reverberation, (b) is with reverberation at T_{60}=200 ms and DRR=11.4 dB, and (c) is T_{60}=1000 ms and DRR=5.76 dB.

amount of reverberation from the word "sail". Under these circumstances, it is no longer possible to determine the original amplitude of the speech, and using amplitude or envelope as a feature will not provide a good estimate. Since amplitude and envelope features do not differentiate different DRR conditions, either different features must be used which enable the reverberation to be distinguished from the speech, or spatial information must be used to localize the source, or otherwise separate the direct signal from the reverberant signal.

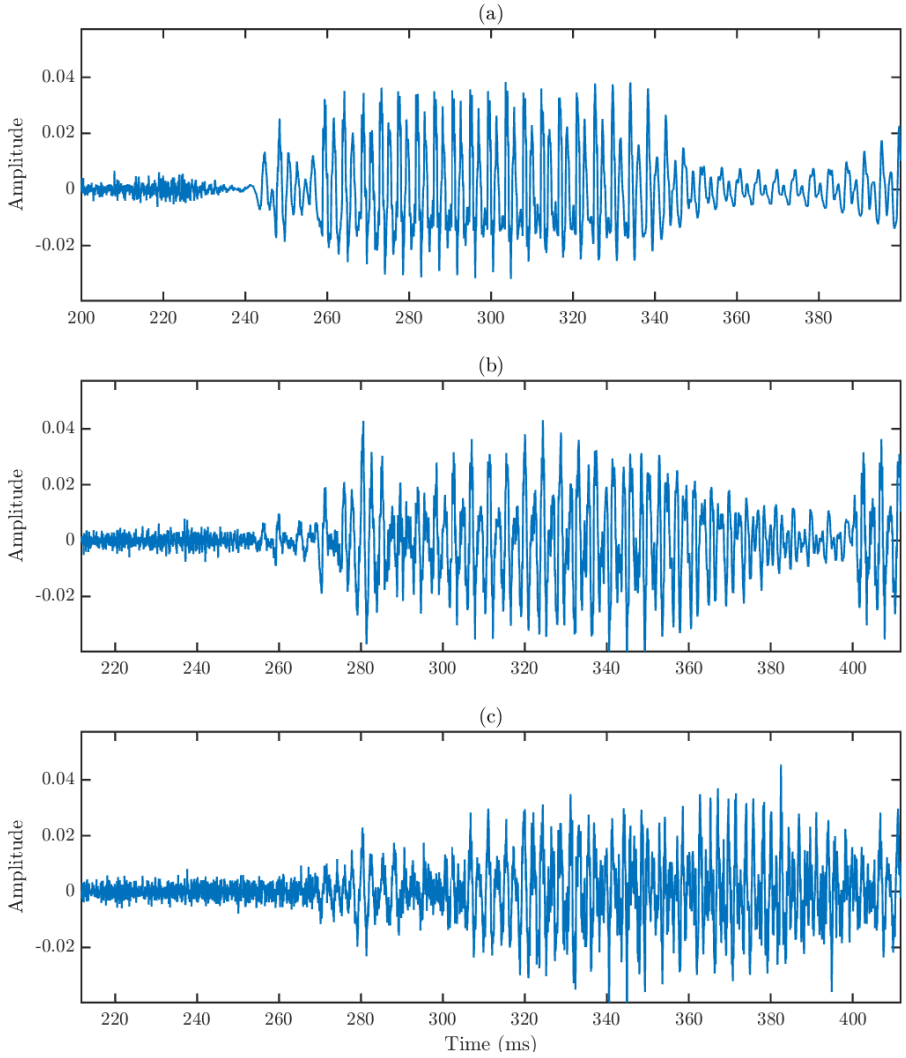

Figure 4.2: TIMIT training dataset utterance "A sailboat may have a bone in her teeth one minute and lie becalmed the next" in low and high T_{60}, low DRR scenarios, (a) without reverberation, (b) with reverberation at T_{60}=200 ms and (c) with reverberation at T_{60}=1000 ms and DRR=−3.79 dB.

4.1.2 Methods for non-intrusive DRR estimation in the literature

Early research into DRR showed that it may be useful in providing distance information in reverberant environments to human listeners as summarised in [116]. Motivated by estimating the distance of a source from the receiver using audio signals, Lu *et al.* [117] took a perceptually motivated approach to non-intrusive DRR estimation based on the human auditory system employing an equalization and cancellation scheme. The equalization used to cancel the direct signal is determined from the instantaneous power ratio of the non-delayed and delayed channels. This equalization scheme is however signal dependent.

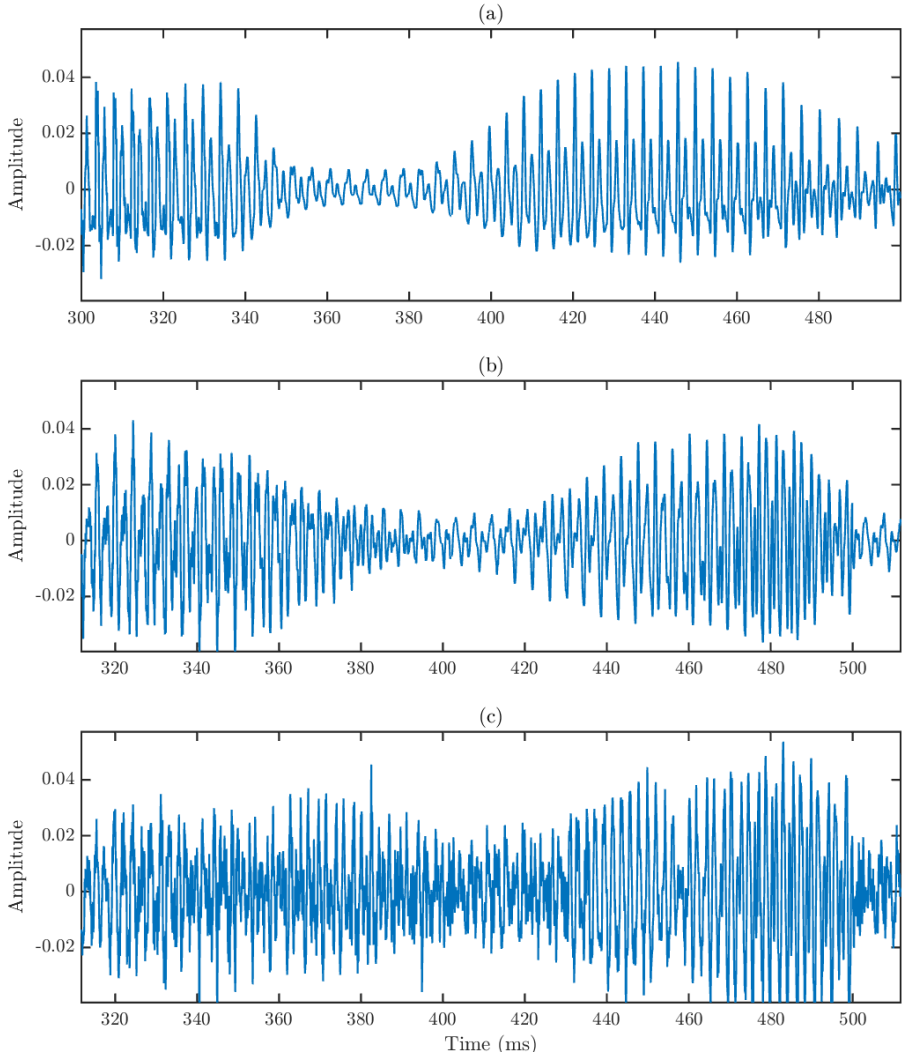

Figure 4.3: Continuation of Fig. 4.2, TIMIT training dataset utterance "A sailboat may have a bone in her teeth one minute and lie becalmed the next" in low and high T_{60}, low DRR scenarios, (a) without reverberation, (b) with reverberation at T_{60}=200 ms and (c) with reverberation at T_{60}=1000 ms and DRR=−3.79 dB.

A correlation between estimated DRR and distance is demonstrated but DRR estimation accuracy is not reported.

Another recent approach to non-intrusive DRR estimation is to use spatial coherence between channels of omnidirectional microphones [118, 119], which was extended to directional microphones [120, 121]. These methods assume that all non-coherent energy is reverberation and that the direct sound is coherent and thus provide the DRR estimate. For negative DRRs, Hioka *et al.*'s method shows errors of approximately 5 dB.

Falk *et al.* [122] observed a correlation between modulation spectrum features and

DRR. Long-term temporal dynamic cues derived from the modulation spectrum produce the SRMR computed in modulation frequency bands and mapped to DRR after training on a suitable speech corpus. Whilst a DRR estimator is proposed, the performance of the DRR estimator is not reported.

In the related task of T_{60} estimation, Dumortier *et al.* [90] proposed using spatial selectivity to enhance the reverberant signal such that its dynamics can be observed more clearly. This is potentially a promising approach for DRR estimation.

4.1.3 Aims and organisation of this chapter

There remains the opportunity to devise an accurate low-complexity non-intrusive DRR estimator from speech. In this chapter, a novel DRR estimation method is proposed where spatial selectivity is used to separate direct and reverberant energy, and account for noise. The method is applicable where the signal was recorded with two or more microphones, as can be found in many mobile communications devices and laptops. The formulation considers the response of the beamformer to reverberant sound and the effect of noise. The method uses a null-steered beamformer to attenuate the direct sound thus yielding the reverberant signal only in the beamformer output. By comparing the power in the direct sound versus the power in the beamformer output, and compensating for the gain of the beamformer, an estimate of the DRR is obtained. The proposed method is thus called DRR Estimation using a Null-Steered Beamformer (DENBE). In simulations, DENBE yields DRR estimates with a standard deviation of 4 dB across a wide variety of room sizes, reverberation times and source-receiver distances. It is also shown that the proposed method is more robust to background noise than a baseline approach. The best estimation accuracy is obtained in the region from -5 to 5 dB which is a relevant range for mobile devices.

The remainder of this chapter is organised as follows: In section 4.2 the DENBE method is presented. In section 4.3 the performance is evaluated, and in section 4.4 the results are compared with the baseline [118]. Finally, conclusions are presented in Section 4.5.

4.2 The DENBE Method

4.2.1 Operating principle

Consider noisy reverberant speech observed in a room. Spatial filtering or beamforming uses a weighted combination of two or more microphone signals to achieve a particular directivity pattern. Consider a beamformer directed at the source of the speech where the beamformer weights are determined such that the source is cancelled out. This is known as a null-steered beamformer. Only the reverberation and noise will appear in the output of the beamformer since the direct signal has been removed. By comparing the noisy reverberant speech signal with the reverberant signal from the output of the beamformer, it may be possible to deduce the clean speech signal and thus the DRR. However, the amount of reverberation in the output of the beamformer emanating from a given direction and frequency depends on the gain of the beamformer in that direction and frequency. Similarly the amount of noise in the output of the beamformer depends on the beamformer gain. By accounting for the gain of the beamformer and the impact of the noise, an accurate estimate of the DRR can be obtained. This is the principle upon which the DENBE method is based.

4.2.2 Acoustic model

This method will be presented first in continuous time and then mapped to the corresponding discrete time for implementation. Also, the formulation is initially presented for an arbitrary number of microphones and does not assume a particular method of solving for the directivity pattern of the beamformer. A reverberant signal $y_m(t)$, continuous in time, t, captured by the m^{th} microphone in an array of M microphones in a room is characterised by the AIR, $h_m(t)$, of the acoustic channel between the source and the m^{th} microphone such that

$$y_m(t) = x'_m(t) + v_m(t), \tag{4.1}$$

where $v_m(t)$ is the additive noise at the m^{th} microphone, and $x'_m(t)$ is the reverberant speech at the m^{th} microphone. The reverberant speech, $x'_m(t)$, can be considered to be a

convolution with the AIR, $h_m(t)$, as

$$x'_m(t) = \int_{\tau=-\infty}^{+\infty} x(\tau) h_m(t-\tau). \tag{4.2}$$

where, $x(t)$, is a speech signal, continuous in time radiating from a given position in a room.

The AIR is a function of the geometry of the room, the reflectivity of the surfaces in the room, and the microphone locations. The AIR for each microphone can be considered in terms of separate AIRs for the direct path, $h_{m,d}(t)$, and the reverberant path, $h_{m,r}(t)$, as

$$h_m(t) = h_{m,d}(t) + h_{m,r}(t), \tag{4.3}$$

The DRR at the m^{th} microphone, $\bar{\gamma}_m$, is the ratio of the power arriving directly at the microphone from the source to power arriving at the m^{th} microphone after being reflected from one or more surfaces in the room [62]. It can be written as

$$\bar{\gamma}_m = \frac{\int |h_{m,d}(t)|^2 dt}{\int |h_{m,r}(t)|^2 dt}. \tag{4.4}$$

When the AIR is convolved with a speech signal, the observation at the m^{th} microphone is the Signal-to-Reverberation Ratio (SRR), Γ, given by

$$\Gamma_m = \frac{E\{|x'_{m,d}(t)|^2\}}{E\{|x'_{m,r}(t)|^2\}}, \tag{4.5}$$

where $E\{\cdot\}$ is the mathematical expectation, the direct speech signal, $x'_{m,d}(t)$, is given by

$$x'_{m,d}(t) = \sum_{\tau=-\infty}^{t} x(\tau) h_{m,d}(t-\tau), \tag{4.6}$$

and $x'_{m,r}(t)$ is defined similarly. The SRR is equal to the DRR in the case when $x(t)$ is spectrally white. The aim of non-intrusive or blind DRR estimation is to estimate γ_m from the observed signals. In the proposed approach spatial selectivity is used to separate the direct and reverberant components of the sound field. Also, the SRR is estimated in narrow frequency bands, thus the result in any one frequency band approaches the DRR.

4.2.3 Beamforming in the frequency domain

The output, $Z(j\omega)$, of a beamformer in the complex frequency domain is given by [123]

$$Z(j\omega) = \mathbf{w}^T(j\omega)\mathbf{y}(j\omega), \tag{4.7}$$

where $\mathbf{w}(j\omega) = [W_0(j\omega), W_1(j\omega), \ldots, W_{M-1}(j\omega)]^T$ is a vector of complex weights for each microphone forming the beamformer response, and $\mathbf{y}(j\omega) = [Y_0(j\omega), Y_1(j\omega), \ldots, Y_{M-1}(j\omega)]^T$ is the vector of microphone signals.

Let the signal at the m^{th} microphone due to a unit plane wave incident on that microphone be $X_m(j\omega, \Omega)$, where $\Omega = (\phi, \theta)$ is the Direction-of-Arrival (DoA), and θ and ϕ are the azimuth and elevation of DoA, respectively. The beam-pattern of the beamformer is given by

$$U(j\omega, \Omega) = \mathbf{w}^T(j\omega)\mathbf{x}(j\omega, \Omega), \tag{4.8}$$

where

$$\mathbf{x}(j\omega, \Omega) = [X_0(j\omega, \Omega), X_1(j\omega, \Omega), \ldots, X_{M-1}(j\omega, \Omega)]^T.$$

4.2.4 Estimation of DRR in the frequency domain

It will now be considered how to use the beamformer to estimate DRR. From (4.1) and (4.3), the signal at the m^{th} microphone in the frequency domain is defined as

$$Y_m(j\omega) = D_m(j\omega) + R_m(j\omega) + V_m(j\omega) \tag{4.9}$$

where $D_m(j\omega) = H_{m,d}(j\omega)S(j\omega)$ is the direct path of the speech at the m^{th} microphone, $R_m(j\omega) = H_{m,r}(j\omega)S(j\omega)$ is the reverberant speech at the m^{th} microphone, and $V(j\omega)$ is the noise at m^{th} microphone. From (4.7),

$$Z_y(j\omega) = Z_d(j\omega) + Z_r(j\omega) + Z_v(j\omega), \tag{4.10}$$

where $Z_d(j\omega) = \mathbf{w}^T(j\omega)\mathbf{d}(j\omega)$ is the beamformer output due to the direct path of the speech signal, $Z_r(j\omega) = \mathbf{w}^T(j\omega)\mathbf{r}(j\omega)$ is the beamformer output due to the reverberant

speech, $Z_v(j\omega) = \mathbf{w}^T(j\omega)\mathbf{v}(j\omega)$ is the noise at the beamfomer output, and where

$$\mathbf{d}(j\omega) = [D_0(j\omega), D_1(j\omega), \ldots, D_{M-1}(j\omega)]^T, \tag{4.11}$$

$$\mathbf{r}(j\omega) = [R_0(j\omega), R_1(j\omega), \ldots, R_{M-1}(j\omega)]^T, \tag{4.12}$$

$$\mathbf{v}(j\omega) = [V_0(j\omega), V_1(j\omega), \ldots, V_{M-1}(j\omega)]^T. \tag{4.13}$$

The beamformer weights, $\mathbf{w}(j\omega)$, are chosen such that $Z_d(j\omega) = 0$, thus

$$Z_y(j\omega) = Z_r(j\omega) + Z_v(j\omega). \tag{4.14}$$

It is assumed that the reverberant energy is the same at all microphones as

$$E\{|R(j\omega)|^2\} = E\{|R_m(j\omega)|^2\} \ \forall m = 1 : M, \tag{4.15}$$

and that the reverberant sound field is isotropic, i.e. it is composed of plane waves arriving from all directions with equal probability and magnitude. The output of the beamformer due to reverberant energy at the microphones can be written in terms of

$$E\{|Z_r(j\omega)|^2\} = G^2(j\omega)E\{|R(j\omega)|^2\}, \tag{4.16}$$

and from (4.8),

$$G^2(j\omega) = \int_\Omega |U(j\omega, \Omega)|^2 \ d\Omega. \tag{4.17}$$

Substituting (4.14) into (4.16) and rearranging, assuming that the reverberation and noise signals are uncorrelated, gives

$$E\{|R(j\omega)|^2\} = \frac{1}{G^2(j\omega)} \left(E\{|Z_y(j\omega)|^2\} - E\{|Z_v(j\omega)|^2\}\right). \tag{4.18}$$

Since it is assumed that the reverberation power is the same at all microphones, from (4.9) and (4.15), then

$$E\{|D_m(j\omega)|^2\} = E\{|Y_m(j\omega)|^2\} - E\{|V_m(j\omega)|^2\} - E\{|R(j\omega)|^2\}. \tag{4.19}$$

The frequency dependent DRR follows from (4.4) as

$$\gamma_m(j\omega) = \frac{E\{|D_m(j\omega)|^2\}}{E\{|R(j\omega)|^2\}}. \tag{4.20}$$

Substituting (4.18) and (4.19) into (4.20) gives

$$\gamma_m(j\omega) = \frac{E\{|Y_m(j\omega)|^2\} - E\{|V_m(j\omega)|^2\}}{\frac{1}{G^2(j\omega)}(E\{|Z_y(j\omega)|^2\} - E\{|Z_v(j\omega)|^2\})} - 1, \tag{4.21}$$

or

$$\gamma_m(j\omega) = \frac{G^2(j\omega)(E\{|Y_m(j\omega)|^2\} - E\{|V_m(j\omega)|^2\})}{E\{|Z_y(j\omega)|^2\} - E\{|Z_v(j\omega)|^2\}} - 1. \tag{4.22}$$

The overall DRR is then given by

$$\bar{\gamma}_m = \frac{1}{\omega_2 - \omega_1} \int_{\omega_1}^{\omega_2} \gamma_m(j\omega)\ d\omega, \tag{4.23}$$

where $\omega_1 \leqslant \omega \leqslant \omega_2$ is the frequency range of interest. This formulation requires knowledge of the noise at the m^{th} microphone, $E\{|V_m(j\omega)|^2\}$, and the noise in the beamformer output, $E\{|Z_v(j\omega)|^2\}$. In a practical application it is assumed that a noise estimator robust to reverberation such as ESG [30] from Chapter 2 can be used to obtain these noise power values.

4.3 Performance evaluation

Three rooms were chosen measuring $\{3\,\text{m}, 4\,\text{m}, 5\,\text{m}\} \times 6\,\text{m} \times 3\,\text{m}$. A two-element microphone array was used with a spacing of 0.062 m to simulate the microphones on an example device such as a laptop. In each room four locations and rotations of the microphone array were chosen at random from a uniform distribution. The source was positioned perpendicular to the microphone array at distances of $0.05, 0.10, 0.50, 1, 2$, and 3 m as shown in Fig. 4.4. No microphone or source was allowed to be less than 0.5 m from any wall. This arrangement was chosen to match the conditions in a real room where the walls are seldom perfectly parallel, and the source-microphone alignment is seldom perfectly parallel with any of the walls. This was performed to avoid large first order reflections from the wall behind the source and the wall behind the microphone directly back to the microphone. AIRs for each room were generated using T_{60} values of 200 to 1000 ms in 100 ms intervals using the

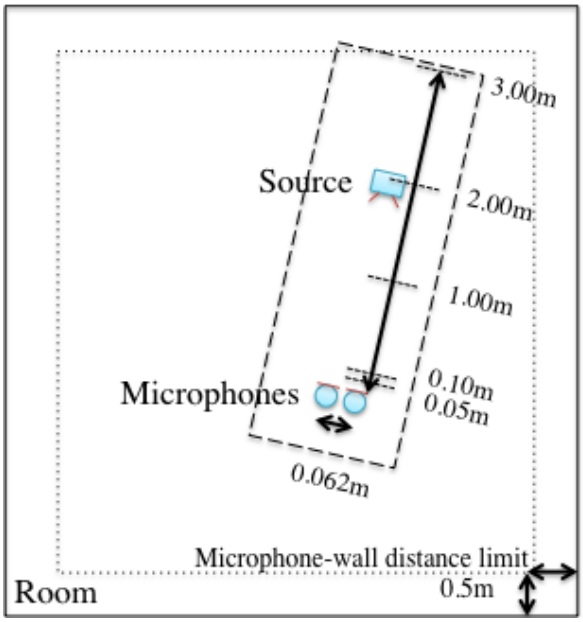

Figure 4.4: Room arrangement for the experiments

source-image method [97, 110]. Speech signals randomly selected from the test partitions of TIMIT [19] were convolved with the AIRs and white noise added at 10, 20, and 30 dB SNR using (3.19). Since the method has potentially broader application than speech since the statistics of speech are not assumed in the method, all experiments were conducted at $f_s = 16\,\text{kHz}$.

Beamformer weights were chosen using a delay-and-subtract scheme to steer a null towards the DoA of the direct path. Since all source positions were equi-distant from the two microphones, this reduces to a simple subtraction giving the familiar dipole beam pattern shown in Fig. 4.5. In practical applications a Time-Difference-of-Arrival (TDoA) estimator such as Generalized Cross-Correlation with Phase Transformation (GCC-PHAT) [124] could be used to set the delay.

Ground truth DRR in fullband was computed for each room, T_{60}, microphone and source position directly from the simulated AIRs. White Gaussian noise was added independently for each microphone at SNRs of 10, 20, and 30 dB where the clean speech power was determined using an implementation of ITU-T P.56 [25, 37].

Two experiments were conducted. In the first experiment the DENBE method using oracle values for the noise powers $E\{|V_m(j\omega)|^2\}$ and $E\{|Z_v(j\omega)|^2\}$ is compared with the

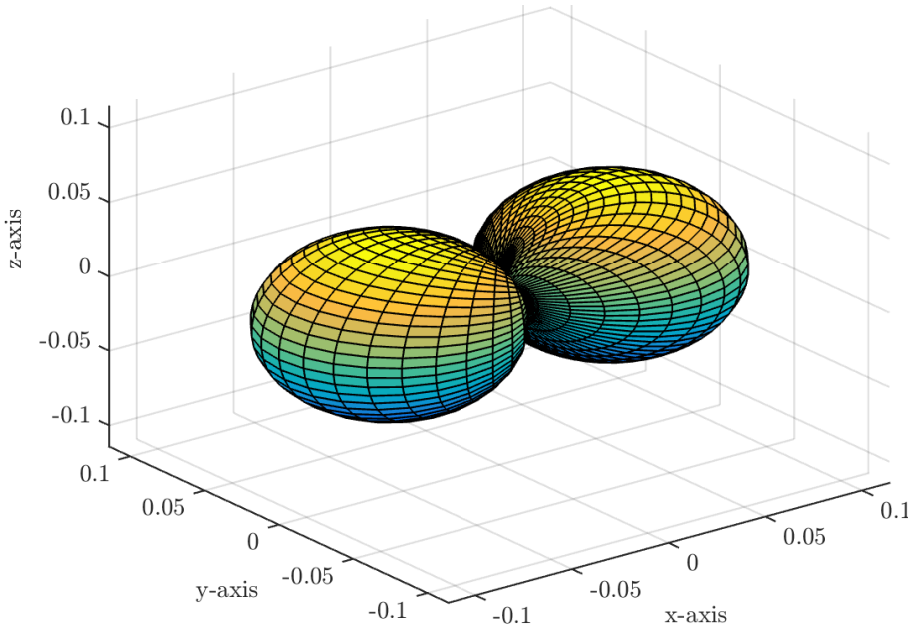

Figure 4.5: 2-channel null-steered beamformer gain and directivity pattern at 200 Hz **with a microphone spacing of** 0.062 m**. The maximum gain is** −9.4 dB**.**

DENBE method where these noise powers are set to zero ignoring the effects of noise, and with the baseline method. In order to evaluate the effects of noise estimation errors on the accuracy of the DRR estimator, a second experiment was conducted with the noise powers $E\{|V_m(j\omega)|^2\}$ and $E\{|Z_v(j\omega)|^2\}$ in (4.22), increased and decreased by 1.5 dB.

The baseline method for comparison was chosen to be that of Jeub *et al.* [118] because it is a recent 2-channel DRR estimation method where the authors claim accurate results. The method returns a vector of estimated DRR by frequency, and the mean of the values $> -\infty$ was used in comparisons.

4.4 Results and discussion

The DRR estimation accuracy of the DENBE and baseline algorithms at SNRs of 10 dB, 20 dB, and 30 dB is shown in Fig. 4.6. To calculate the first and second order statistics, DRR estimates were binned according to their ground truth DRR in 1 dB intervals.

As can be seen in Fig. 4.6, the DENBE method estimates with a standard deviation of 4 dB across DRRs ranging from −5 to 5 dB. As DRR decreases the DENBE method tends to

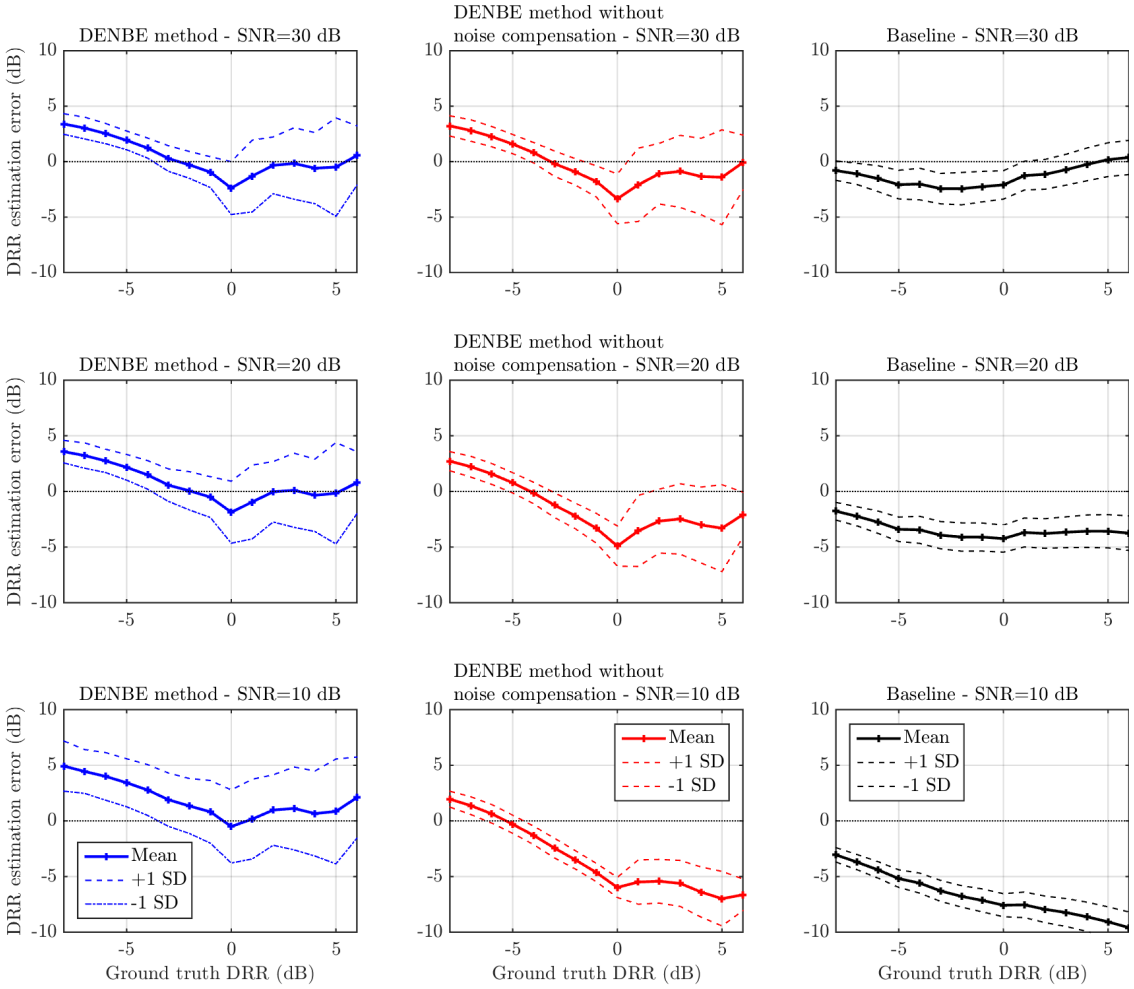

Figure 4.6: Results for the DENBE method, DENBE method without noise compensation, and the baseline at $10, 20$ **and** $30\,\text{dB}$ **SNR. Results have been aggregated into** $1\,\text{dB}$ **bins**

overestimate DRR. This is a result of the assumption that reflections arrive from all angles with equal probability. For a particular room and T_{60}, lower DRRs are obtained with larger source-microphone distances. This in turn results in the strong early reflections arriving from directions which are closer to the direct path DoA and are therefore more attenuated by the beamformer null. By under-accounting for these early reflections in (4.16) the DRR is overestimated.

For positive DRRs the mean estimate is unbiased but has a relatively high variance. Again this is most likely due to the assumption that the reverberation is isotropic. However, the distribution of strong early reflections is more random than for negative DRRs, and so they will be attenuated to a greater or lesser extent by the beam pattern depending on

their direction of arrival.

The importance of including noise in the formulation is shown by comparing the DENBE method with and without noise compensation to the baseline. Without noise compensation the DENBE method follows the tendency of the baseline to underestimate DRR as noise increases. Conversely with the DENBE method (with noise included in the formulation) accuracy is consistent across the range of SNRs shown with an increase of approximately $\pm$ 1 dB in the standard deviation of the estimates at SNR = 10 dB.

The sensitivity to errors in noise estimation at the reference microphone and at the output of the beamformer is shown in Fig. 4.7. Where there are errors of opposite polarity affecting the direct and beamformed power, the DRR estimates remain close to the case where there is no error, effectively cancelling each other out. Where the errors are of the same polarity, there is an additive effect with a $\pm$1.5 dB error on each term leading to a $\pm$3 dB error overall. This suggests that the method is more sensitive to the bias in a noise estimator than its variance.

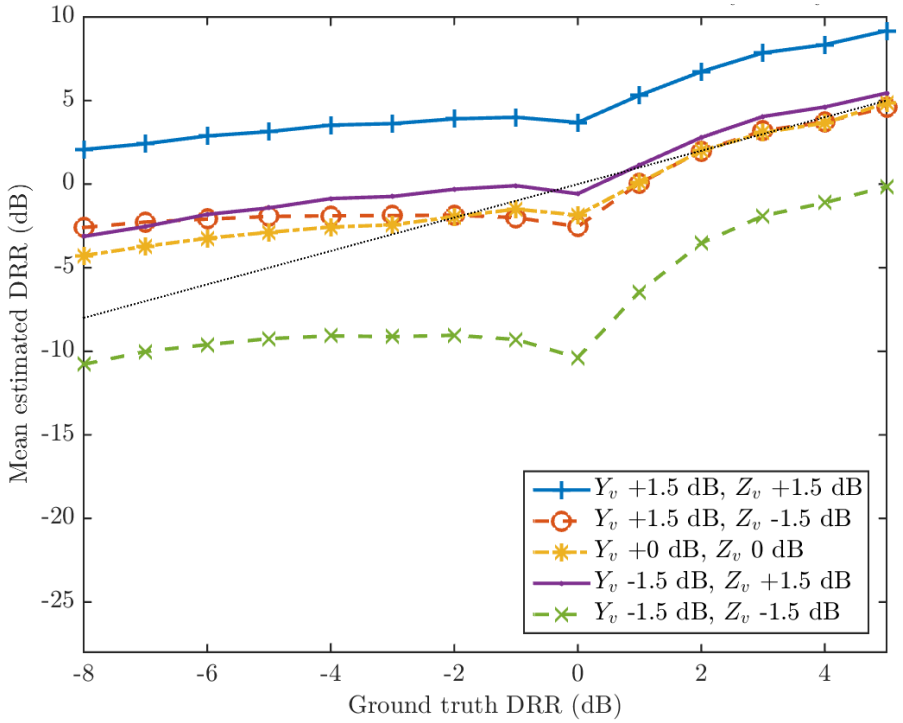

Figure 4.7: Effect of noise estimation errors on mean DRR estimates at 20 dB SNR

4.5 Conclusion

A novel method for non-intrusive estimation of DRR from multi-channel speech taking noise into account has been presented. The method shows more robustness to noise at realistic SNRs than a baseline based on spatial coherence. This is achieved by compensating for the bias caused by the presence of uncorrelated noise at the microphones. Whilst the tests performed were limited to two channel configurations, the method can be applied to a multi-channel system with an arbitrary number of microphones with the selection of an appropriate beamformer.

The formulation returns an estimate of DRR as a function of frequency, and therefore a frequency dependent DRR could be provided by transforming the estimates by STFT frequency bin into for example International Organization for Standardization (ISO) preferred frequencies [59] which has potential application in efficient dereverberation schemes. The capability to non-intrusively estimate DRR in ISO frequency bands will be further explored in Chapter 6. Also, since the method does not rely on the statistics of speech it could also be applied to music.

Chapter 5

Non-intrusive codec-robust clipping detection for degraded speech

5.1 Introduction

CLIPPING is a common problem in voice telecommunications. Clipping of speech occurs when the amplitude of a signal exceeds the available dynamic range. Clipping in speech telecommunications is undesirable because it reduces the subjective quality [125], may result in a loss of intelligibility, and degrades the performance of subsequent speech processing such as ASR or dereverberation [62]. Whilst mobile handset manufacturers have developed technologies to handle very high amplitude signals up to 120 dB Sound Pressure Level (SPL) [126], the use of low cost microphones, and high levels of ambient noise [22], contribute to a high likelihood of the occurrence of clipping. The application in [126] suggests that sound pressure might exceed the capability of a typical handset microphone by a factor of 20 dB, and it is shown in [127] that the speech level from a single talker may vary by as much as 30 dB.

Codecs are widely used in telecommunication networks and speech recording devices, and in general, it is not known by a clipping detection algorithm whether the speech has passed through a codec or not. In many practical cases, such as in the receiver side in

mobile cellular telephony and video conferencing, the speech prior to passing through a codec is not available. *Waveform codecs* such as G.711 [128] aim to preserve waveform shape and thus features of clipping after decoding. To minimize the bandwidth required for transmission, codecs used in mobile communications typically divide the sound into frames, and then attempt to create a similar sound to the current frame that minimizes the error, or cost function, between the observed sound and the created sound, using methods such as Code-excited Linear Prediction (CELP) [129]. However, *perceptual codecs* such as Global System For Mobile Communications (GSM) 06.10 [130] and Adaptive Multi-Rate (AMR) [131] modify the cost function to take account of the perceptual significance of errors so that they no longer minimize waveform errors. Perceptual codecs do not aim to preserve waveform shape after decoding [132], and therefore will not preserve features of clipping after decoding as will be demonstrated in Section 5.1.2.

There is therefore a need to be able to detect, and where possible correct, the effects of clipping in noisy speech that has passed through a codec in order to restore the original signal.

5.1.1 Definitions of clipped and decoded speech

A clipped sampled speech signal is given by

$$\tilde{x}(n) = \begin{cases} \eta_+\mu_+, & \text{if } x(n) > \eta_+\mu_+ \\ x(n), & \text{if } \eta_-\mu_- \leqslant x(n) \leqslant \eta_+\mu_+ \\ \eta_-\mu_-, & \text{if } x(n) < \eta_-\mu_- \end{cases}. \tag{5.1}$$

where $x(n)$ is the original speech signal, $\tilde{x}(n)$ is the clipped speech signal, $\mu_+ = \max(x(n))$ and $\mu_- = \min(x(n))$, the amplitudes of the most positive and negative samples in $x(n)$ respectively, and η_+ and η_- are the clipping levels for positive and negative excursions respectively, and n is the sample index. A signal whose amplitude is multiplied by a factor of 2 and clipped at the original peak absolute amplitude has a $\eta = 0.5$. Clipping level is also defined relative to 0 dBFS such that

$$\eta_{dBFS} = 20\log_{10}(\eta). \tag{5.2}$$

There is a wide range of standards used to determine headroom in sound programme connections [133–136]. For the purposes of this work, typical clipping levels are assumed to lie in the range of $\eta = 0.1$ to $\eta = 1$.

The term *decoded speech* is used in this paper to refer to speech that has been encoded and subsequently decoded through combinations of one or more waveform or perceptual codecs. *Null codec* is used to refer to a codec which passes the signal unchanged. *unencoded speech* is used to refer to speech that has not passed through a codec, or has passed through the null codec.

5.1.2 Properties of clipped speech

Properties of clipped speech in the time domain

The properties of clipped speech will now be analysed in the time and frequency domains. First the amplitude histogram will be used to analyse clipping in the time domain. The amplitude pdf of speech has been modelled as a gamma distribution with a shaping parameter between 0.4 and 0.5 [137, 138], and also as a Laplacian distribution [137, 139]. The log amplitude histogram of an unclipped speech utterance will therefore typically form an approximately triangular distribution as will be shown below.

Figure 5.1 (a)–(d) shows: the unclipped time domain waveforms, (e)–(h), their normalized log amplitude histograms, and (i)–(l), their spectrograms for utterance "Will Robin wear a yellow lilly" from the test dataset of the TIMIT [19] speech corpus. The utterance was down-sampled from 16 kHz to 8 kHz, normalized to ± 1 amplitude, and then coded-decoded using codecs listed in Table 5.1.

Name	Type
Null	Passes the signal unchanged
G.711 [128]	Waveform codec
GSM 06.10 [130]	Perceptual codec
AMR [131] at 4.75 kbps	Perceptual codec

Table 5.1: Codecs used in Figs. 5.1 and 5.2

The log amplitude histograms in Fig. 5.1 (e)–(f) have a triangular shape that varies depending on the codec applied. In the case of G.711 this is a result of A-law non-linear

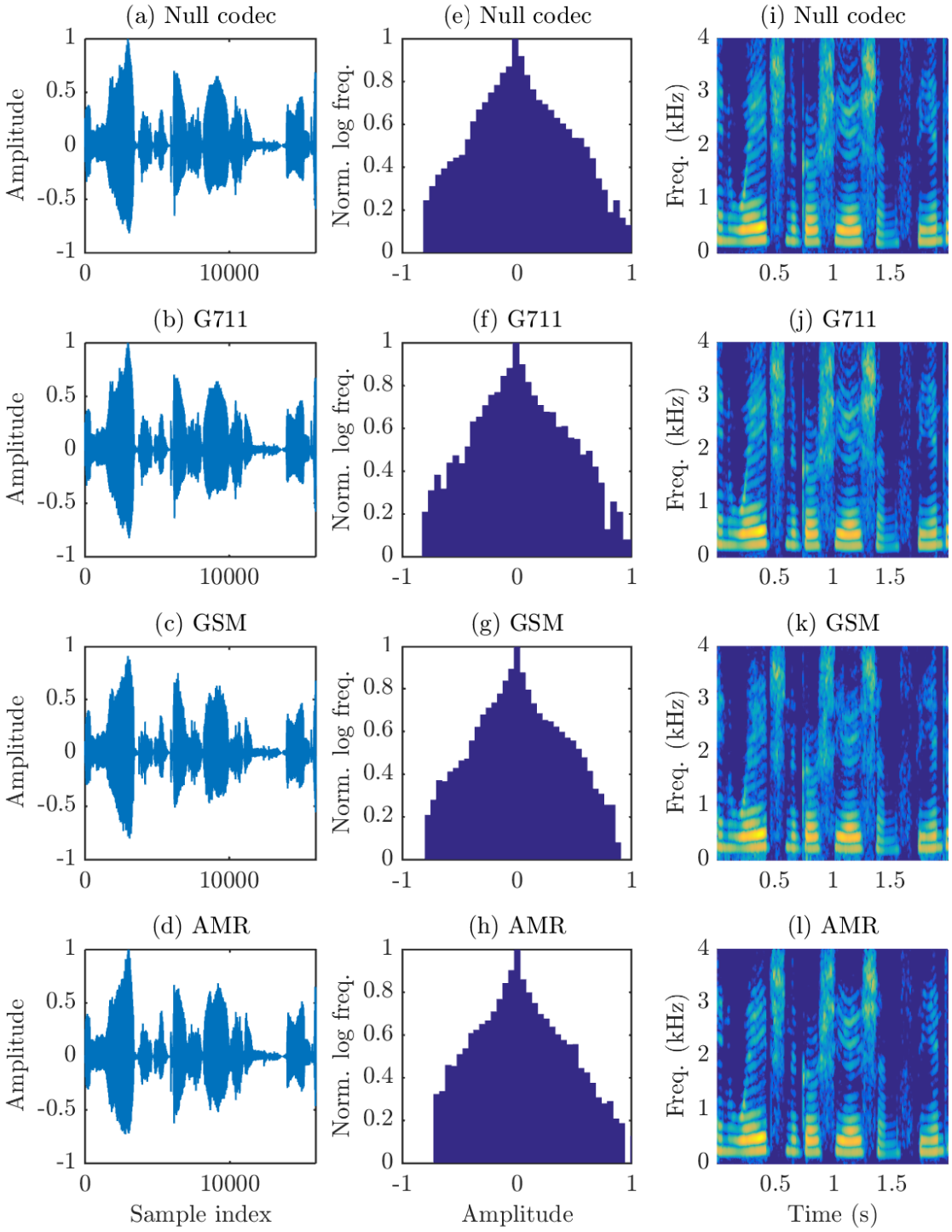

Figure 5.1: Waveforms, log amplitude histograms, and spectrograms for the unclipped signals (η =0 dBFS**) for** TIMIT **test utterance "Will Robin wear a yellow lilly"**

quantization. There are small differences in the non-linear quantization of μ-law (used in North America and Japan) decoded speech versus from A-law (used in Europe), but they

are not significant in relation to this illustration. In the case of GSM, and AMR it is due to the perceptual coding and decoding process.

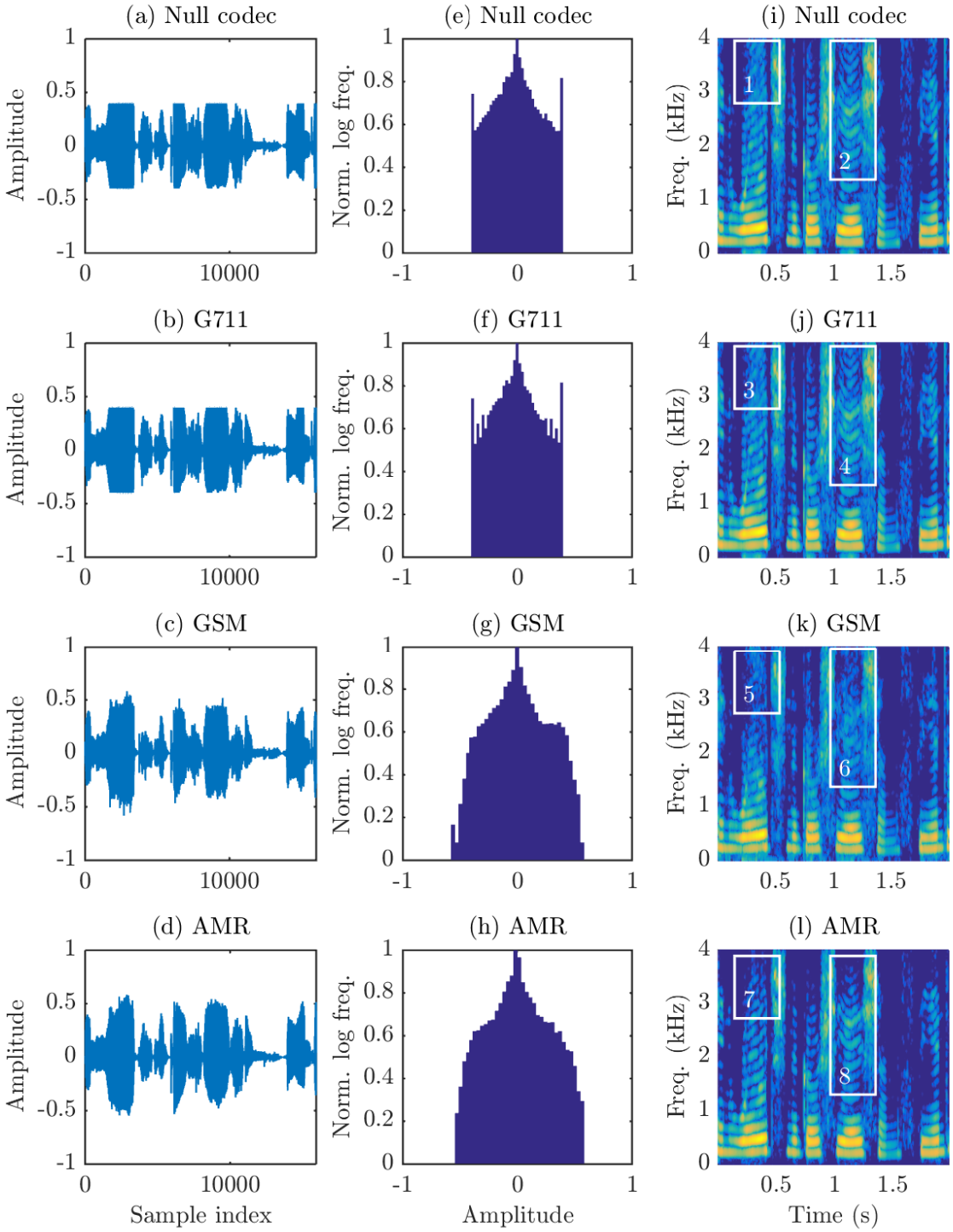

Figure 5.2: Waveforms, log amplitude histograms, and spectrograms for the signals clipped at $\eta = -8$ dBFS

Figure 5.2 shows the TIMIT utterance "Will Robin wear a yellow lilly" clipped according to (5.1) at $\eta = -8$ dBFS, and processed through each of the codecs listed in Table 5.1. The characteristic limiting of the speech waveform peaks is clearly visible in Figs. 5.2 (a)–(b) when compared with Figs. 5.1 (a)–(b). However, with the perceptual codecs in (c)–(d), although the signals appear somewhat clipped, the peaks no longer have a rigid maximum because perceptual codecs alter signal phase, and cannot encode the non-linearities caused by clipping.

For the unencoded and waveform-coded clipped speech as shown in Figs. 5.2 (e)–(f), when compared to the unclipped speech in Figs. 5.1 (e)–(f), the main body of the histogram is preserved. However, samples with amplitudes exceeding the clipping level appear at the clipping level, since the amplitude of the speech signal cannot exceed the clipping level. This creates a spike in the histogram at the clipping level for the positive and negative excursions of the clipped signal at the clipping level. For the perceptually decoded speech shown in Figs. 5.2 (g)–(h), the body of the histograms are broadened and their tails are shortened, but maintain a similar overall shape to the unclipped utterance histograms. Here, amplitude values at the clipping level have been spread out across a range of values by the codec, thus removing the spike observable in the unencoded signal in Figs. 5.2 (e)–(f).

Properties of clipped speech in the frequency domain

The frequency domain properties of clipping will now be analysed using the spectrogram. Clipping a signal generates additional frequencies not present in the original signal [140]. Roughness, or fluctuations of amplitude that occur due to constructive and destructive interference of sound waves, is a measure of the dissonance between adjacent harmonics of a complex tone [141]. A clipped speech signal, $\tilde{x}(n)$, is expected to exhibit more roughness than an unclipped speech signal because new frequencies have been introduced by the clipping process. Also, processing speech with a perceptual codec itself introduces distortion [142], as well as band-limiting the speech signal.

The spectrograms for unclipped and clipped speech for the utterance "Will Robin wear a yellow lilly" are shown alongside the log amplitude histograms in Figs. 5.1 and Figs. 5.2. In the spectrograms, lighter colours indicate where there is more energy in

the signal. Horizontal lines show the harmonics of the signal. Significant areas of the spectrograms are annotated from 1 to 8.

In clipped speech, there is more energy in the upper parts of the spectrogram when compared to the same unclipped speech as can be seen in Figs. 5.2 (i)–(j) and Figs. 5.1 (i)–(j). The parts of the clipped speech signals indicated by annotations 1-4 have more energy in highest harmonics than in the unclipped signal. The harmonics of the clipped speech are at a similar level across all frequencies, whereas in the unclipped speech signal, the harmonics are at different levels. As the clipping level decreases so that more samples are clipped, these effects are accentuated. After perceptual decoding, the effects of clipping as described above are less pronounced, as indicated by annotations 5-8.

It can be concluded that in both the time and frequency domain, features of clipping in decoded speech are obscured by the perceptual decoding process. Therefore clipping detection in perceptually decoded speech is a more difficult than in unencoded speech because the codecs used to compress the speech data alter the features of clipped speech.

5.1.3 Clipping detection and restoration algorithms in the literature

Clipping restoration is usually a two stage process involving first detection of the clipped samples, and then subsequent restoration of the clipped samples. Clipping detection and restoration algorithms in the literature fall into the following broad categories:

1. Methods that identify when a large number of samples fall within a narrow band of values at the extremities of the signal [143–148];
2. Methods that identify features of clipping in the amplitude histogram [149];
3. Methods that identify unexpected non-linearities in the speech corresponding to clipping [150];
4. Methods which use linear prediction to predict the unclipped signal from the clipped signal and remove unexpected deviations [151];
5. Methods that operate in the frequency domain and predict the original spectrum [152].
6. Methods which assume prior knowledge of the clipped samples [18, 153–155] and therefore do not deal with the problem of detection.

The method for detecting clipped samples in a clipped speech signal used in [143–148], relies upon the existence of flattened peaks in the clipped signal. This method of detection is described in Atlas *et al.* [148], is repeated here since it is widely used, and will be used later in this chapter as a baseline and in combination with other methods.

Let $\tilde{x}(n)$ be $x(n)$ of length N samples after clipping according to (5.1). The set of sample indices $\hat{\mathbf{c}}$ at which samples in $\tilde{x}(n)$ are clipped is defined as

$$\hat{\mathbf{c}} = \{n : 0 \leqslant n < N \text{ and } (\tilde{x}(n) > \psi_+ \text{or } \tilde{x}(n) < \psi_-)\} \tag{5.3}$$

where

$$\psi_+ = (1 - \epsilon) \max_n(\tilde{x}(n)) \tag{5.4}$$

and

$$\psi_- = (1 - \epsilon) \min_n(\tilde{x}(n)), \tag{5.5}$$

for some tolerance ϵ, such as 0.01. As can be seen from Fig. 5.2 (g)–(h) when compared with (e)–(f), the amplitude extremities of the log amplitude histogram of perceptually decoded clipped speech lie within a much larger range than for unencoded and waveform decoded speech. Therefore this method will require $\epsilon \gg 0.01$ to identify clipped samples in perceptually decoded speech. In increasing ϵ, it will be necessary to make a trade-off between many missed detections or many false alarms.

The methods of [149] use features of the time-domain amplitude histogram, $h(i)$, of unencoded speech to estimate clipping, where i is the index of each histogram bin. In the first embodiment of [149], a reference amplitude histogram template is subtracted from the normalized amplitude histogram, $h(i)$, of the clipped speech signal, $\tilde{x}(n)$, and differences between the two are used to determine the presence or absence of clipping.

In [150], the derivative of the signal is used to detect the onset and end-points of clipped sets of samples for unencoded speech, and in [151], linear prediction is used to restore the estimated original signal for unencoded speech.

In [152], a spectral transformation of each frequency band using a model spectral envelope is proposed for unencoded speech. The authors observe that clipping introduces new peaks into the speech spectrum in the lower frequencies, and that in higher frequencies, the spectral envelope is flattened. A Gaussian Mixture Model (GMM) approach is used to

model the peaks in the lower frequencies of the speech spectrum and weight the largest two peaks assumed to be the first and second formants higher than the rest of the spectrum. In the higher frequencies, a clean spectral envelope based on prior knowledge of the phoneme is used to compensate for the flattening effect of clipping. The amount of compensation to apply is controlled using an estimate of the clipping ratio [156].

The clipping detection and restoration schemes in the literature attempt to solve the problem of detecting clipped samples in clean speech. Detecting clipped samples in speech that has passed through a perceptual codec is relatively new and no literature was identified which attempted to solve this problem.

5.1.4 Aims and organisation of this chapter

This chapter seeks to address the problem of detecting hard-clipping in both unencoded, waveform decoded and perceptually decoded speech without requiring prior knowledge of the unclipped signal or the codec used. Motivated by mobile applications, the aim is to produce reliable algorithms for identifying the clipped samples in a clipped decoded speech signal suitable for realtime use, that could form part of overall restoration schemes.

The contributions of this chapter are

1. to present an analysis, derivation and evaluation of an time-, frequency- and hybrid time- and frequency-domain algorithms for non-intruslvely detecting clipping;
2. to show how these algorithms are robust to perceptual codecs and noise; and
3. to compare the results of the proposed algorithms with a clipping detector from the literature [148] optimized to decoded speech.

The remainder of this chapter is organised as follows: In Section 5.2 a time domain clipping detection algorithm is presented. In Section 5.3, a frequency domain clipping detection algorithm is presented. In Section 5.4, a hybrid time- and frequency-domain clipping detection algorithm is presented based on Iterated Logarithm Amplitude Histogram (ILAH) and Least Squares Residuals (LSR). In Section 5.5, an improved hybrid time- and frequency-domain clipping detection algorithm is presented. In Section 5.6, an improved and simplified time domain detection algorithm is proposed based on ILAH. In Section 5.7,

the test approach is described along with the method for detecting clipping which is used as a baseline. In Section 5.8, the outcomes of the tests conducted are reported, and in Section 5.9, conclusions are drawn.

5.2 ILAH clipping detection algorithm

5.2.1 Operating principles of the ILAH algorithm

The Iterated Logarithm Amplitude Histogram (ILAH) clipping detection algorithm is a time-domain histogram-based algorithm for estimating clipping in a speech signal which has been encoded and subsequently decoded. An assumption of the ILAH algorithm is that the pdf of the speech utterance being analysed approximates the long-term pdf of speech as modelled in [137–139], and that the amplitude histogram approximates the pdf.

The ILAH algorithm transforms the shape of the amplitude histogram by taking successive logarithms of the histogram frequencies. This results in an approximately pentagonal shaped envelope which can be modelled with a series of straight lines to simplify analysis and to deal with variations between talkers and utterances. Figure 5.3 shows an illustrative ILAH of clipped speech together with straight line approximations to the histogram envelope labeled (a) to (d). The triangular-shaped spread of values beneath

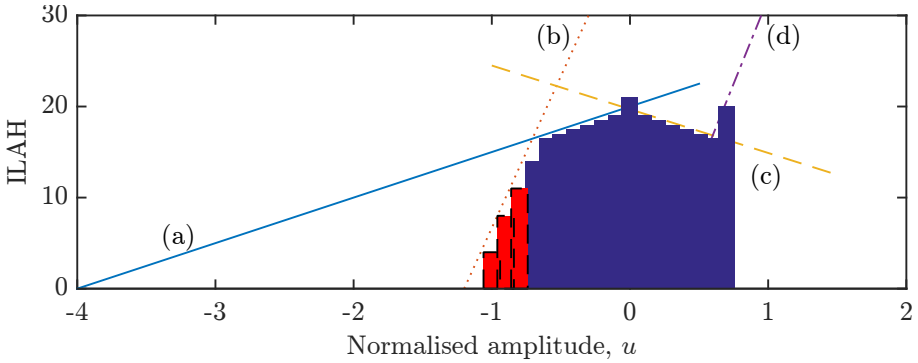

Figure 5.3: An illustrative ILAH showing the a clipped, perceptually decoded speech signal on the negative side and a nul-coded signal on the positive side. The red bars with dashed outline below (b) are caused by the perceptual decoding process.

letter (b) for the negative-amplitudes is assumed also to spread inwards towards the centre of the histogram, forming a skirt-like shape. In the unencoded or waveform decoded case, an estimate of the outer gradient in the vicinity of the negative and positive extremities

of the histogram using only the non-zero values will be positive on the positive side as shown at (d), and negative on the negative side. In the decoded case, where no spikes are present, the gradient will be negative on the positive side and positive on the negative side as shown at (b). The ILAH is typically asymmetric.

The negative and positive sides of the histogram each typically has an inner gradient near the centre of the histogram and an outer gradient near the extremities of the histogram. The inner gradients (a) and (c) have been found to be related to the original signal level, whilst the outer gradient for example (b) has been found to be related to the type of codec used. The iterated logarithm, denoted $\log\log(\cdot)$, is used to produce the transformed histogram, $\tilde{h}(i)$ from the amplitude histogram $h(i)$, where i is the histogram bin index. As was discussed in Section 5.1.2, the log amplitude histogram of speech produces and approximately triangular distribution. Taking a second logarithm and setting $\tilde{h}(i) = 0$ where $\tilde{h}(i) \leqslant 0$ truncates the sides of the triangle, and flattens the inner gradients, thus producing a pentagonal-shaped histogram. This was verified experimentally for a wide range of speech utterances.

Using the shape of the histogram, the ϵ parameter from (5.4) and (5.5) is estimated for the speech utterance, and then used in (5.3) to estimate $\hat{\mathbf{c}}$, the clipping level, η, and the unclipped peak signal amplitude, u_o.

To illustrate the distinctive features of the transformed amplitude histogram of clipped speech, $\tilde{h}(i)$, the utterance "Even then, if she took one step forward he could catch her" from the TIMIT [19] training dataset was normalized to satisfy $\max(|x(n)|) = 1$, and passed through different codecs including the null codec, both unclipped and clipped at $\eta = -10.5\,\text{dBFS}$. Figure 5.4 (a) to (d) show $\tilde{h}(i)$ for the unclipped speech, whilst Fig. 5.4 (e) to (f) show $\tilde{h}(i)$ for the clipped speech.

Clipped samples in $\tilde{x}(n)$ cause the negative and positive extremities of $\tilde{h}(i)$ to be truncated resulting in an approximately pentagonal-shaped histogram as illustrated in Figs. 5.4 (e) to (g). In clipped speech after waveform coding, a spike at the amplitude extremities is observable in $\tilde{h}(i)$, whereas after perceptually decoding, the histogram sides are no longer vertical, and are wider at the base of the histogram at the amplitude extremities as shown in Figs. 5.4 (g)–(h).

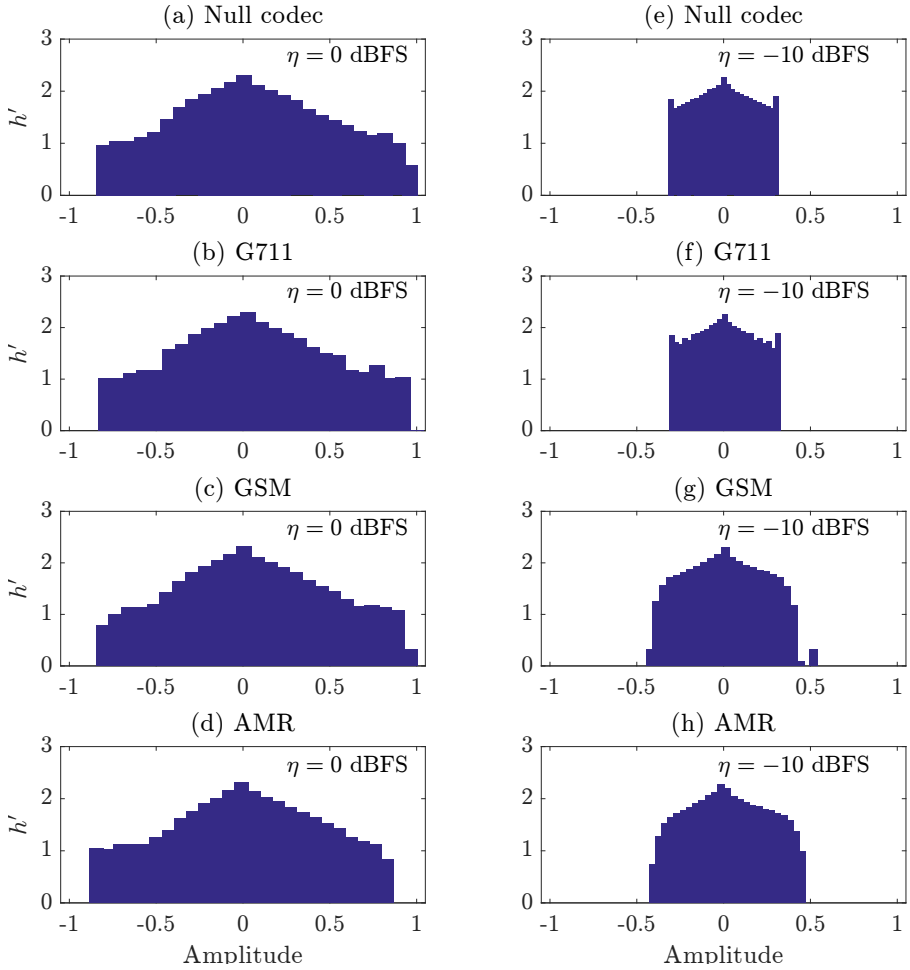

Figure 5.4: ILAHs of TIMIT training dataset utterance "Even then, if she took one step forward he could catch her" for each codec, normalized with no clipping (plots (a) to (d)) and clipped at $\eta = -10$ dBFS (plots (e) to (h)).

5.2.2 ILAH algorithm description

ILAH transform

To perform the ILAH transform, first any DC offset from the signal is first removed using mean subtraction. The signal is then normalised as

$$\bar{\tilde{x}}(n) = (\tilde{x}(n) - \mu_{\tilde{x}(n)})/\zeta \tag{5.6}$$

where the normalization factor, $\zeta = \max(|\tilde{x}(n) - \mu_{\tilde{x}(n)}|)$, and $\mu_{\tilde{x}(n)}$ is the mean of the clipped signal. The ILAH is determined by generating the I-bin amplitude histogram, $h(i)$, and $u(i)$, from normalised mean-subtracted clipped speech, $\bar{\tilde{x}}(n)$. Positive and negative

sides of the signal, denoted by the $_+$ and $_-$ subscripts, are processed independently within the algorithm. The ILAH, $\tilde{h}(i)$, is then computed as

$$\tilde{h}(i) = \begin{cases} \log\log\{h(i)\}, & \text{if } h(i) > 1 \\ 0, & \text{otherwise.} \end{cases} \tag{5.7}$$

A continuous speech pdf is asymptotic to zero at the maximal positive and negative amplitudes. With a histogram however, the smallest non-zero value of $\tilde{h}(i_{max})$ and $\tilde{h}(i_{min})$ is 1. Therefore, to ensure that this asymptotic behaviour is well approximated $N \gg I$.

ILAH gradient and original signal level estimation

To determine the gradients of the positive and negative inner sides of the histogram, an optimal straight line, shown in Fig. 5.3 (a) and (c), is fitted using least squares [157] and extended to the point where it intersects $h(i) \to 0$, and similarly for the positive case (c). Let $ai + b = f(U, H)$ represent a first order least squares fit of the values $H = [h_0, h_1, \ldots, h_{L-1}]$ over the interval spanned by $I = [i_0, i_1, \ldots, i_{L-1}]$ where $U = u(I)$. Let i_s and i_m denote offsets for the inner and outer extremities of each side of the histogram. The inner gradients of $\tilde{h}(i)$ are found as follows:

$$\begin{aligned} a_+ i + b_+ &= f([u(i_{0+} + i_s) \ldots u(i_{max} - i_m)], \\ &\quad\ [\tilde{h}(i_{0+} + i_s) \ldots \tilde{h}(i_{max} - i_m)]) \\ a_- i + b_- &= f(u(i_{min} + i_m) \ldots u(i_{0-} - i_s), \\ &\quad\ [\tilde{h}(i_{min} + i_m) \ldots \tilde{h}(i_{0-} - i_s)]). \end{aligned} \tag{5.8}$$

To ensure that the gradient estimates, a_+ and a_-, result in a finite estimates for $\hat{u}_{o+}$ and $\hat{u}_{o-}$, a_+ and a_- are limited to a suitable minimum value, ψ. This limit is applied as follows:

$$\begin{aligned} a_+ &= \begin{cases} a_+, & \text{if } a_+ \leqslant \psi \\ -\psi, & \text{otherwise} \end{cases} \\ a_- &= \begin{cases} a_-, & \text{if } a_- \geqslant \psi \\ \psi, & \text{otherwise.} \end{cases} \end{aligned} \tag{5.9}$$

The zero-crossing points of the inner gradients provide an estimate of the original signal level before clipping. If the zero-crossing point is sufficiently close to the extremities of the histogram, the signal will be deemed to be unclipped. However, these trajectories show a bias. Therefore a third order mapping function, $\tilde{g}(\cdot)$ trained on the ground truth original peak signal amplitudes is used to compensate. The training dataset is described in Section 5.7.1. Estimates for the positive and negative peak original signal levels are given by

$$\hat{u}_{o+} = \tilde{g}(-\zeta b_+/a_+ + \mu_{\tilde{x}(n)}), \tag{5.10}$$

and

$$\hat{u}_{o-} = \tilde{g}(-\zeta b_-/a_- + \mu_{\tilde{x}(n)}), \tag{5.11}$$

and ζ is the normalization factor defined in (5.6).

ILAH clipping level estimation

Estimation of clipping level, η, is next considered. The estimation method is dependent on whether or not the speech has been processed by a perceptual codec. As discussed in Section 5.2.1, the outer gradients of the ILAH give an indication of whether a perceptual codec was applied to the speech signal. The gradients in the vicinity of $u(i_{max})$ and $u(i_{min})$ respectively must therefore be estimated. A first order least squares fit is performed on the positive and negative outer sides of the histogram using the last i_m non-zero histogram values as follows:

$$\begin{aligned} d_+ i + e_+ &= f([u(i_{max} - i_m) \dots u(i_{max}), \\ &\qquad \tilde{h}(i_{max} - i_m) \dots \tilde{h}(i_{max})]) \\ d_- i + e_- &= f([u(i_{min}) \dots u(i_{min} + i_m), \\ &\qquad \tilde{h}(i_{min}) \dots \tilde{h}(i_{min} + i_m)]). \end{aligned} \tag{5.12}$$

This method will not perform well as $\eta \to 0\,\text{dBFS}$, because fluctuations at the extremities of the histogram could lead to large gradient estimation errors. To inhibit this mechanism for $\eta \to 0\,\text{dBFS}$, a threshold, β, is therefore defined as shown in (5.13) and (5.14) to prevent estimation close to 0 dBFS. β is chosen to minimise η estimation error over the training clipped speech dataset over all codecs described in Section 5.7.1.

To allow for variations in the shape of the histogram when estimating whether or

not the signal has passed through a perceptual codec, a further threshold, γ, is used. The value of γ is chosen to ensure that a perceptual codec is only detected when the outer gradients ILAH slope away significantly downwards from the inner gradients of the ILAH, thus helping prevent small changes in the gradient of unclipped signals from being identified as clipped. Therefore when the following condition is true, a perceptual codec has been detected for positive excursions

$$\hat{P}_+ = \begin{cases} 1, & \text{if } (d_+ + \gamma < a_+) \text{ and } (u(i_{max})/\hat{u}_{o+} < \beta) \\ 0, & \text{otherwise,} \end{cases} \tag{5.13}$$

and similarly for negative excursions:

$$\hat{P}_- = \begin{cases} 1, & \text{if } (d_- > a_- + \gamma) \text{ and } (u(i_{min})/\hat{u}_{o-} > \beta) \\ 0, & \text{otherwise.} \end{cases} \tag{5.14}$$

where $\hat{P}_+$ and $\hat{P}_-$ are Boolean variables which are true when a perceptual codec has been detected on each respective side of the ILAH. The overall determination of whether a perceptual codec has been detected is as follows:

$$\hat{P} = \hat{P}_+ \text{ or } \hat{P}_- \tag{5.15}$$

The inner bounds of the spread of values caused by the perceptual codec is first estimated as follows:

$$\hat{u}_{p+min} = 2\hat{u}_{p+} - u(i_{max}), \tag{5.16}$$

and similarly for negative excursions:

$$\hat{u}_{p-min} = 2\hat{u}_{p-} - u(i_{min}), \tag{5.17}$$

where $\hat{u}_{p+min}$ and $\hat{u}_{p-min}$ are the inner bounds of the spread of values, and the intersections of the lines on each side in (5.8) and (5.12), $\hat{u}_{p+}$ and $\hat{u}_{p-}$ respectively, are found as follows:

$$\begin{aligned} \hat{u}_{p+} &= (a_+ - d_+)/(e_+ - b_+) \\ \hat{u}_{p-} &= (a_- - d_-)/(e_- - b_-). \end{aligned} \tag{5.18}$$

To control how much of the spread of values produced by the codec is included when estimating η, a constant, the skirt factor, θ, determined experimentally, is used as follows:

$$\begin{aligned} \hat{u}_{p+min} &= 2\theta\hat{u}_{p+} - u(i_{max}) \\ \hat{u}_{p-min} &= 2\theta\hat{u}_{p-} - u(i_{min}). \end{aligned} \tag{5.19}$$

When a perceptual codec has been detected, η is estimated by comparing the inner bounds to the estimated peak absolute amplitude of the signal as follows:

$$\begin{aligned} \hat{\eta}_{p+} &= \hat{u}_{p+min}/\hat{u}_{o+} \\ \hat{\eta}_{p-} &= \hat{u}_{p+min}/\hat{u}_{o-}. \end{aligned} \tag{5.20}$$

When no perceptual codec has been detected, the extremal excursions are used as follows:

$$\begin{aligned} \hat{\eta}_{+} &= u(i_{max})/\hat{u}_{o+} \\ \hat{\eta}_{-} &= u(i_{min})/\hat{u}_{o-}. \end{aligned} \tag{5.21}$$

Clipping vector estimation

The vector of estimated clipped samples, $\hat{\mathbf{c}}$, is now be estimated. Equation (5.3) is used to estimate which samples in $x(n)$ are clipped by setting $\epsilon = 1 - \hat{\eta}_+$ and $\epsilon = \hat{\eta}_- - 1$ for positive and negative excursions respectively. In order to ensure samples close to this level are classified as clipped, a further constant, ϵ_n, is added to ϵ for non-encoded speech, or ϵ_p for decoded speech determined experimentally. The estimate of which samples are clipped is determined as follows:

$$\hat{c}(n) = \begin{cases} 1, & \text{if } \{(x(n)/\hat{u}_{o+} > \hat{\eta}_+ - \epsilon_n) \text{ and not } \hat{P}_+\} \\ & \quad \text{or } \{(x(n)/\hat{u}_{o-} < \hat{\eta}_- + \epsilon_n) \text{ and not } \hat{P}_-\} \\ & \quad \text{or } \{(x(n)/\hat{u}_{o+} > \hat{\eta}_{p+} - \epsilon_p) \text{ and } \hat{P}_+\} \\ & \quad \text{or } \{(x(n)/\hat{u}_{o-} < \hat{\eta}_{p-} + \epsilon_p) \text{ and } \hat{P}_-\} \\ 0, & \text{otherwise.} \end{cases} \tag{5.22}$$

ILAH algorithm summary

The ILAH algorithm can therefore be summarised as follows:

Inputs: $\tilde{x}(n)$.

Outputs: $\hat{\mathbf{c}}$, $\hat{\eta}_+$, $\hat{\eta}_-$, $\hat{\eta}_{p+}$, $\hat{\eta}_{p-}$, $\widehat{C}$, $\hat{P}$, $\hat{u}_{o+}$, $\hat{u}_{o-}$.

1) Normalise $\tilde{x}(n)$ and then compute $\tilde{h}(i)$ from (5.7).

2) Estimate the gradients of left and right top of main body of $\log\log\{h(i)\}$ from (5.8), applying the minimum limits for small gradients from (5.9) if necessary.

3) Estimate peak absolute amplitudes from (5.10) and (5.11).

4) Estimate gradients of extremities of histogram from (5.12).

5) Test for a perceptual codec from (5.13) and (5.14), and if detected compute perceptual thresholds from (5.19), and $\hat{\eta}_{p+}$ and $\hat{\eta}_{p-}$ from (5.20), otherwise compute $\hat{\eta}_+$ and $\hat{\eta}_-$ from (5.21).

6) If $\widehat{C}$ is false, determine clipped samples, $\hat{c}$, using (5.22), otherwise set $\hat{c}$ to zeroes.

Results for the ILAH algorithm are presented in Section 5.8.

5.3 The LSR clipping detection algorithm

The PSD of a speech signal, $x(n)$, of length N samples will typically decrease with frequency and can be represented using the Long Term Average Speech Spectrum (LTASS) [158]. Clipping affects the roughness of the signal spectrum as discussed in Section 5.1, and is expected to cause the spectrum to deviate from LTASS. On the assumption that within the bandwidth of speech, roughness is not significantly altered by the coding-decoding process, by fitting a straight line to the PSD in the STFT domain using least squares, and comparing the deviation, an estimate of the clipping is obtained. This is the hypothesis upon which the Least Squares Residuals (LSR) clipping detection algorithm is based.

5.3.1 LSR algorithm description

Let $S(l,k)$ be the PSD of $\tilde{x}(n)$ in the STFT domain at timeframe l, and frequency index k. Then let

$$\mathbf{S}_l = [S(l,0)\ S(l,1)\ \ldots\ S(l,K-1)]^T, \tag{5.23}$$

and let $\hat{\mathbf{S}}_l$ be a first order least squares estimate of $\mathbf{S}_l$ for frame l, where K is the number of frequency bins in each frame, and there are L frames in total. Also, let the residual, $r(l)$, be the RMS difference between $\hat{\mathbf{S}}_l$, and $\mathbf{S}_l$, defined as

$$r(l) = \sqrt{\sum_{k=0}^{K-1} \left(\hat{S}(l,k) - S(l,k)\right)^2} \tag{5.24}$$

In an unclipped speech signal, it is expected that $\hat{\mathbf{S}}_l$ will follow $\mathbf{S}_l$ closely, resulting in a small $r(l)$. However, in frames where the speech is clipped, the roughness of the spectrum is expected to result in a larger $r(l)$. In order to determine $\mathbf{S}_l$ in the vicinity of each sample, a Hamming window of length 0.25 ms zero-extended to 2 ms, and an overlap of 75% is used for the STFT. The frequency bins used are restricted to those representing the main speech energy bands up from 0 to 1.5 kHz. In order to derive the least squares estimate of $\hat{\mathbf{S}}_l$, let

$$r(l) \quad = \quad \|(a + b\mathbf{k}) - \mathbf{S}_l\|_2 \,, \tag{5.25}$$

where $r(l)$ is the residual of the least squares regression which is to be minimized, $(a, b) \in \mathbb{R}$, and $\mathbf{k} = [1 \ldots K]$. Reformulating to produce an expression in the form $\mathbf{Ax} = \mathbf{b}$ [159] gives

$$r(l) = \|\mathbf{A}\mathbf{x}_l - \mathbf{S}_l\|_2 \,, \tag{5.26}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & K \end{bmatrix}^T \text{ and } \mathbf{x}_l = \begin{bmatrix} a \\ b \end{bmatrix}. \tag{5.27}$$

To solve for $\mathbf{x}_l$ setting $r(l) = 0$ gives

$$\mathbf{x}_l = \mathbf{A}^\dagger \mathbf{S}_l, \tag{5.28}$$

where $\mathbf{A}^\dagger$ is the Moore-Penrose pseudo inverse, $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ [160], which is employed since $\mathbf{A}$ is not square. Since (5.26) requires the computation of $\mathbf{A}\mathbf{x}_l - \mathbf{S}_l$ in the vicinity of every sample, a further step is made to avoid repeated computation. Writing

$$\mathbf{A}\mathbf{x}_l - \mathbf{S}_l = \mathbf{A}\mathbf{A}^\dagger \mathbf{S}_l - \mathbf{S}_l, \tag{5.29}$$

and then reformulating gives

$$\mathbf{A}\mathbf{x}_l - \mathbf{S}_l = \mathbf{Q}\mathbf{S}_l, \tag{5.30}$$

where

$$\mathbf{Q} = \mathbf{A}\mathbf{A}^{\dagger} - \mathbf{I}, \tag{5.31}$$

and $\mathbf{A}^{\dagger}$ represents the Moore-Penrose inverse of $\mathbf{A}$ [160]. Substituting (5.30) into (5.26) gives

$$r(l) = \|\mathbf{Q}\mathbf{S}_l\|_2 . \tag{5.32}$$

It is only necessary to compute $\mathbf{Q}$ once for LSR, and it can be statically defined since it is only dependent on the value of K.

To determine whether a frame contains a clipped sample, a vector, $\mathbf{r}$ is created containing the residual computation from each frame such that

$$\mathbf{r} = [r(0)\ r(1)\ \ldots\ r(L-1)]. \tag{5.33}$$

This is then normalized over all L frames in the signal as

$$\bar{\mathbf{r}} = \mathbf{r}/\max(\mathbf{r}). \tag{5.34}$$

In order to determine which samples are clipped, the L-based residual estimate, $\bar{\mathbf{r}}$ must be aligned with the time domain samples. Since there is an overlap factor of 75%, there are 3 fewer frames than samples. The frame-based normalized residual estimate, $\bar{\mathbf{r}}$, is aligned with the sample index, n, and zero-extended as

$$\bar{r}_n(n) = \begin{cases} \bar{r}(n-1), & \text{where } 0 < n < L-2 \\ 0, & \text{otherwise,} \end{cases} \tag{5.35}$$

where $\bar{r}_n(n)$ is the normalized residual estimate based on the sample index, n. This alignment, as a result of using the STFT with a 75% overlap factor does however mean that if the first or last two samples are clipped, they will not be detected. A vector indicating

the presence of clipped samples can therefore be estimated as

$$\hat{c}(n) = \begin{cases} 1, & \text{where } \bar{r}_n(n) > \rho \\ 0, & \text{otherwise,} \end{cases} \tag{5.36}$$

where ρ is a threshold above which it is assumed that a sample is clipped. The clipping vector in the case of LSR, $\hat{\mathbf{c}}$ is defined as

$$\hat{\mathbf{c}} = [\hat{c}(0)\ \hat{c}(1)\ \ldots\ \hat{c}(N-1)]. \tag{5.37}$$

The optimum threshold, ρ_o, is determined by estimating $\hat{\mathbf{c}}$ on a training corpus of speech at different clipping levels and codecs, using a range of ρ, and selecting the value of ρ which gives an Equal Error Rate (EER) [161] for the detection of clipped samples.

LSR algorithm summary

The LSR algorithm can be summarised as follows:

Inputs: $\tilde{x}(n)$, f_s; Outputs: $\hat{\mathbf{c}}$.

1) Normalize the clipped speech signal, $\tilde{x}(n)$.

2) Compute the periodogram, $\mathbf{S}_l$, in short frames.

3) Fit a straight line, $\hat{\mathbf{S}}_l$, to the frequency bands up to 3.3 kHz for each frame using least squares from (5.26).

4) Compute the residuals, $\mathbf{r}$, from (5.24), then normalize the residuals vector $\mathbf{r}$ to give $\bar{\mathbf{r}}$ from (5.34).

5) Determine clipped samples based on residuals in each frame exceeding ρ_o from (5.35) and (5.36). Results for the LSR algorithm are presented in Section 5.8.

5.4 The LILAH clipping detection algorithm

A drawback of LSR is that it has relatively high complexity compared to ILAH, and also does not achieve similar results to ILAH in unencoded clipped speech. The LSR-ILAH (LILAH)

algorithm is an unsupervised hybrid approach that aims to obtain the best results of both ILAH and LSR, whilst reducing overall computational complexity.

To obtain the best estimation results of both algorithms, the perceptual codec detection flags, $\hat{P}_+$ and $\hat{P}_-$, are determined by the ILAH algorithm and used to switch from the ILAH to the LSR estimate of $\hat{\mathbf{c}}$ if a perceptual codec is detected. To reduce the computational complexity of LSR, ILAH is used to estimate zones, z where there is a high likelihood that clipping has occurred such that LSR need only be performed within these zones. A clipping zone, $z(n_1, n_2)$, where n_1 and n_2 are the start and end sample indices of the zone respectively, is defined as a range of samples where clipped samples within that zone are separated by not more than τ seconds. The vector of Clipping zones, $\hat{\mathbf{z}}$, is estimated based on the output of ILAH. A sparse speech signal containing non-zero samples only in the clipping zones, $\bar{\mathbf{x}}$, is obtained as follows:

$$\bar{\mathbf{x}} = \mathbf{x}^T \hat{\mathbf{z}}, \tag{5.38}$$

The resultant signal, $\bar{\mathbf{x}}$, is then processed using LSR as before. Results for the LILAH algorithm are presented in Section 5.8.

5.5 The CRACD clipping detection algorithm

Codec-robust Automatic Clipping Detector (CRACD) is an unsupervised hybrid approach based on LILAH which uses the estimated unclipped peak signal levels, $\hat{u}_{o+}$ and $\hat{u}_{o-}$, obtained as described in Section 5.2 from ILAH to modify LSR to improve estimation accuracy in heavily clipped signals, i.e. where η is low.

In a heavily clipped signal, as η decreases and the signal becomes more distorted, the normalization in (5.34) of LSR will be skewed by the increasing number of clipped samples. Therefore one side of the estimated clipping level, e.g. $\hat{\eta}_+$ from (5.20) is used to control ρ at low η, thus

$$\rho = \begin{cases} \rho_o \hat{\eta}_+, & \text{if } \hat{\eta}_+ < \eta_n \\ \rho_o, & \text{otherwise} \end{cases} \tag{5.39}$$

where η_n is the normalisation switching threshold, determined experimentally, and set, in

this work, to $\eta_n = -10.5\,\text{dBFS}$. The improvement over LILAH is dependent on the accuracy of the estimate for $\hat{u}_{o+}$ and $\hat{u}_{o-}$. Poor estimates of $\hat{u}_{o+}$ and $\hat{u}_{o-}$ will reduce the accuracy of CRACD at $\eta \to 1$ in comparison with LILAH.

CRACD algorithm summary

The CRACD algorithm is summarised as follows:

Inputs: $\tilde{x}(n)$, f_s.

Outputs: $\hat{\mathbf{c}}$, $\hat{\eta}_+$, $\hat{\eta}_-$, $\hat{\eta}_{p+}$, $\hat{\eta}_{p-}$, $\widehat{C}$, $\hat{P}$, $\hat{u}_{o+}$, $\hat{u}_{o-}$.

1) Perform ILAH on $\tilde{x}(n)$ to obtain clipping estimates $\hat{\mathbf{c}}$, $\hat{\eta}_+$, $\hat{\eta}_-$, $\hat{\eta}_{p+}$, $\hat{\eta}_{p-}$, $\widehat{C}$, $\hat{P}$.

2) If $\hat{P}$ (perceptual codec detected) then

   a) Estimate the clipping zones, $\hat{\mathbf{z}}$ from $\hat{\mathbf{c}}$ from ILAH.

   b) Compute $\bar{\mathbf{x}}$ from (5.38).

   c) Compute ρ from (5.39).

   d) Using $\bar{\mathbf{x}}$ and f_s as input, perform LSR as defined in Section 5.3 using ρ from 2c above.

   e) Return $\hat{\mathbf{c}}$ from the results of LSR instead of ILAH.

Results for the CRACD algorithm are presented in Section 5.8.

5.6 The ILAH2 clipping detection algorithm

An improved Iterated Logarithm Amplitude Histogram (ILAH2) algorithm, is now presented. The motivation for developing this algorithm arises because both ILAH and LSR rely on several parameters which must be trained individually and successively using a speech corpus making the algorithms impractical. In contrast, the ILAH2 algorithm either dispenses with the parameters needed by the previously presented algorithms or determines them automatically. The only remaining parameters required are not critical in the selection of values and originate from prior art. Only a single training step is needed to map the estimated original signal level.

As discussed in Section 5.1.4, no prior knowledge of whether the signal has passed through a codec is assumed. In the case where no codec or waveform coding has been applied existing algorithms can be used to determine the clipping level. If this case can be identified from the clipped speech, the overall detection accuracy will increase. Therefore in the ILAH2 algorithm, the case where the speech signal is clipped and has not passed through a perceptual codec is first identified using a method based on [148]. In a clipped unencoded signal, a significant number of samples will have amplitude values in a small range close to the maximum and minimum amplitudes of the signal. By determining the ratio of the samples within this range to the total number of samples in the signal, an estimate for whether the signal is clipped and unencoded is found. Therefore if the ratio of samples above $(1-\epsilon)\max_n(\tilde{x}(n))$ or below $(1-\epsilon)\min_n(\tilde{x}(n))$ to the signal length exceeds a threshold, ϕ_c, the signal is deemed to be clipped and not perceptually decoded, thus

$$\phi = \left(\left(\sum_{n:\tilde{x}(n) \geqslant (1-\epsilon)\max_n(\tilde{x}(n))} 1 \right) + \left(\sum_{n:\tilde{x}(n) \leqslant (1-\epsilon)\min_n(\tilde{x}(n))} 1 \right) \right) / N \tag{5.40}$$

where ϕ is the ratio of the number of samples of the amplitude of $\tilde{x}(n)$ within ϵ of the maximum and minimum amplitudes, ϕ, to the total number of samples in the utterance.

The purpose of ϵ is to capture the range of digital values resulting from the conversion of the analogue full-scale speech signal amplitude value. The value of ϕ_c chosen of 0.001 was found experimentally to produce good results in combination with $\epsilon = 0.001$ using clipped TIMIT speech samples over η from 0.1 to 1. This value is not critical since the ILAH analysis will also identify clipped non-perceptually encoded cases. The clipped and unencoded case flag, C, is then estimated as follows:

$$\widehat{C} \quad = \quad \begin{cases} 1, & \text{if } \phi \geqslant \phi_c \\ 0, & \text{otherwise.} \end{cases} \tag{5.41}$$

where ϕ_c is the minimum number of samples at the maximum or minimum amplitude for the utterance to be considered clipped and not having passed through a perceptual codec.

The signal is then analysed using the ILAH transform from Section 5.2.2. A series of steps are required to determine the inner and outer gradients of the positive side of the transformed histogram. Any DC offset is removed using mean subtraction and the signal

is normalised as

$$\bar{\tilde{x}}(n) = (\tilde{x}(n) - \mu_{\tilde{x}(n)})/\zeta \tag{5.42}$$

where the normalization factor, $\zeta = \max(|\tilde{x}(n)|)$, and $\mu_{\tilde{x}(n)}$ is the mean of the clipped utterance, $\tilde{x}(n)$. Let $h(i)$ denote the histogram of $\bar{\tilde{x}}(n)$, where i is the index of each histogram bin. The signal amplitude at bin centre i is given by

$$u(i) = \min(\bar{\tilde{x}}(n)) + \left(i - \frac{1}{2}\right) \frac{\max(\bar{\tilde{x}}(n)) - \min(\bar{\tilde{x}}(n))}{I} \tag{5.43}$$

where $i : 0 \leqslant i < I - 1$, I is the number of histogram bins. In order to maintain the shape of the histogram independent of utterance length, I is determined using Doane's formula [162], as

$$I = 1 + \log_2(N) + \log_2\left(1 + \frac{|g1|}{\sigma_{g1}}\right) \tag{5.44}$$

where N is the length of the clipped utterance, $\tilde{x}(n)$, $g1$ is the skewness of the distribution of the speech, and

$$\sigma_{g1} = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}}. \tag{5.45}$$

The histogram is then transformed using the ILAH transform as in 5.7.

The inner and outer gradients of the positive side of $\tilde{h}(i)$ are then found by treating the positive side of the histogram envelope as a piecewise linear least squares problem with two line segments. The point of intersection of the line segments is assumed to lie at the centre of one of the histogram bins between the smallest positive bar in the histogram, i_{0+}, and the largest positive bar, i_{max}. A search is performed to fit a straight line to the points either side of a point of intersection with the lowest combined residual. This yields an intersection location, χ, an inner line of the form, $y = a_+x + b_+$, and an outer line, $y = d_+x + e_+$. The positive and negative clipping levels, η_+ and η_- respectively, are defined as follows

$$\eta_+ = \begin{cases} (1-\epsilon), & \text{if } \widehat{C}, \\ (u(i_{max})(1-\epsilon) + \mu_{\tilde{x}(n)})/\zeta, & \text{if } d_+ > 0, \\ (2u(\chi_+) - u(i_{max}) + \mu_{\tilde{x}(n)})/\zeta, & d_+ < a_+, \\ 1, & \text{otherwise.} \end{cases} \tag{5.46}$$

$$\eta_- = \begin{cases} (\epsilon - 1), & \text{if } \widehat{C}, \\ (u(i_{min})(1-\epsilon) + \mu_{\tilde{x}(n)})/\zeta, & \text{if } d_+ > 0, \\ (2u(\chi_-) - u(i_{min}) + \mu_{\tilde{x}(n)})/\zeta, & d_+ < a_+, \\ 1, & \text{otherwise,} \end{cases} \tag{5.47}$$

where the inner and outer extremities of the left and right sides of the histogram are defined as

$$\begin{aligned} i_{max} &= \arg\max_i\{\tilde{h}(i) > 0\}, \\ i_{min} &= \arg\min_i\{\tilde{h}(i) > 0\}, \\ i_{0+} &= \arg\min_i\{u(i) > 0\}, \\ i_{0-} &= \arg\max_i\{u(i) < 0\}. \end{aligned} \tag{5.48}$$

As discussed in Section 5.2.1, an unencoded or waveform coded clipped signal is expected to have a peak in the histogram at the amplitude extremities which is absent for perceptually decoded signals. The presence of this peak can be determined by comparing the outer and inner gradients. If no peak is present, and the signal is clipped, the positive outer gradient, d_+, is expected to be less than the positive inner gradient, a_+. The robustness of this test is increased by only indicating that a perceptual codec is present if the signal was determined to not be clipped and unencoded as indicated by $\widehat{C}$. Detection of a perceptual codec, indicated by $\hat{P}$, is therefore given by

$$\hat{P} = \begin{cases} 1, & \text{if } d_+ < a_+ \text{ and not } \widehat{C}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.49}$$

The clipping vector, $\hat{\mathbf{c}}$ is computed as

$$\hat{\mathbf{c}} = \{n : 0 \leqslant n < N \text{ and } (\tilde{x}(n)/\zeta > \eta_+ \text{or } \tilde{x}(n)/\zeta < \eta_-)\}. \tag{5.50}$$

The positive side of the histogram was found to provide sufficient information to determine the clipping state of the signal. From (5.46) and (5.47) it can also be seen that when the outer positive gradient is greater than the inner positive gradient and less than zero, the signal is deemed not to be clipped.

The zero-crossing points of the inner gradients provide an estimate of the original

signal level before clipping. However, these trajectories show a bias. Bias compensation for the negative and positive estimates is performed using mapping functions $g_+(\cdot)$, and $g_-(\cdot)$, trained trained on a suitable training corpus using the actual and estimated peak original signal amplitudes. Estimates for the positive and negative peak original signal levels are then given by

$$\hat{u}_{o+} = g_+(-\zeta b_+/a_+ + \mu_{\tilde{x}(n)}), \tag{5.51}$$

and

$$\hat{u}_{o-} = g_-(-\zeta b_-/a_- + \mu_{\tilde{x}(n)}). \tag{5.52}$$

ILAH2 algorithm summary

The ILAH2 algorithm can therefore be summarised as follows:

Inputs: $\tilde{x}(n)$.

Outputs: $\hat{\mathbf{c}}$, $\hat{\eta}_+$, $\hat{\eta}_-$, $\widehat{C}$, $\hat{P}$, $\hat{u}_{o+}$, $\hat{u}_{o-}$.

1) Test $\tilde{x}(n)$ for the case where the signal is clipped and unencoded thus estimating $\widehat{C}$ from (5.41) and compute $\hat{\mathbf{c}}$ from (5.3).

2) Remove DC offset, normalise $\tilde{x}(n)$ and then compute $\tilde{h}(i)$ from (5.7) using I determined from (5.44).

3) Determine the positive and negative gradient intersection locations, χ_+ and χ_- and thus positive and negative clipping levels η_+, and η_- respectively from (5.46) and (5.47).

4) Test for a perceptual codec using (5.49).

5) If $\widehat{C}$ is false, determine clipped samples, $\hat{c}$, using (5.50), otherwise set $\hat{c}$ to zeroes.

6) Estimate peak absolute amplitudes from (5.51) and (5.52).

Results for the ILAH2 algorithm are presented in Section 5.8.

5.7 Performance evaluation

5.7.1 Training

For the purposes of training $g_+(\cdot)$ and $g_+(\cdot)$, a training clipped speech dataset was defined. A set of 24 male and 24 female speech files were randomly selected from the TIMIT training dataset and downsampled to 8 kHz. Clipping levels of −20, −16, −14, −12, −10, −8, −5, −3, -2, -1, and 0 dBFS were used to represent a system where sound pressure exceeds design limits by up to 20 dB [126]. Babble noise from NOISEX-92 [20] was downsampled to 8 kHz and mixed with the clean speech at SNRs of 0 to 20 dB in 5 dB steps prior to clipping. Codecs G.711 [128], GSM 06.10 [130], and AMR-Narrowband [131] at 12.2 kbit/s were applied after mixing with noise. An SNR of 20 dB has been assumed as a suitable baseline case for telecommunications applications.

5.7.2 Testing

For testing, the same dataset configuration was used excepting that speech files were selected from the TIMIT test dataset using different talkers. The SA1 and SA2 sentences were excluded from the datasets so that no common text was used.

All of the proposed algorithms, ILAH, LSR, LILAH, CRACD, and ILAH2 are compared with a baseline at 11 clipping levels and with four codecs using the Receiver Operating Characteristic (ROC) [161] to analyse the results. The algorithms are evaluated in terms of mean Accuracy (ACC), ACC variance by utterance, mean False Positive Rate (FPR), and mean F1. The measures are computed as follows:

$$\begin{aligned} ACC &= (\mathcal{TP} + \mathcal{TN})/(\mathcal{P} + \mathcal{N}) \\ F1 &= 2\mathcal{TP}/(2\mathcal{TP} + \mathcal{FP} + \mathcal{FN}), \\ \mathcal{FPR} &= \mathcal{FP}/\mathcal{N} \end{aligned} \tag{5.53}$$

where $\mathcal{TP}$ is the number of True Positives, (samples that are clipped correctly identified as clipped); $\mathcal{TN}$ is the number of True Negatives (samples that are not clipped correctly identified as unclipped); $\mathcal{FP}$ is the number of False Positives (samples that are not clipped incorrectly identified as clipped); $\mathcal{FN}$ is the number of False Negatives (samples that are clipped identified as unclipped); $\mathcal{P}$ is the total number of samples identified as clipped,

both correctly and incorrectly, and $\mathcal{N}$ is the total number of samples identified as unclipped, both correctly and incorrectly.

Mean ACC is a measure of how accurately the algorithms identify clipped and unclipped samples as a proportion of the total number of samples.

Mean F1 is a measure of the correlation between $\hat{\mathbf{c}}$ for each algorithm and the ground truth, and is a guide to overall performance. ACC variance is a measure of how well each algorithm adapts to the differences between utterances and talkers. A low variance means that the algorithm ACC is utterance independent.

The $\mathcal{FPR}$ is a measure of the proportion of unclipped samples incorrectly identified as clipped. Where the detection results are to be used for LP-based restoration algorithms, False Negative Rate (FNR) should also be considered.

5.7.3 Algorithm set-up used for comparison

ILAH algorithm

The following parameters for the ILAH algorithm determined experimentally for best F1 were used: $\epsilon_w = 0.001, I = 25, i_s = 2.i_m = 3, \psi = 0.005, \beta = 0.9, \gamma = 2, \theta = 1.0569, \epsilon_p = 0.009$, and $\epsilon_n = 0.001$, and natural logarithms were used in the ILAH histogram.

LSR algorithm

The following parameters for the LSR and LILAH algorithm determined experimentally for best F1 were used: $\rho_o = 0.3963$,

LILAH and CRACD

The LILAH and CRACD algorithms use the same parameters as ILAH and LSR.

ILAH2 algorithm

The ILAH2 algorithm used the following parameter: $\epsilon_n = 0.001$.

Optimized baseline algorithm

The clipping detection method of [148] using (5.3) discussed in Section 5.1 above will now be optimized to the context of this chapter and then used as a baseline to evaluate the performance of the algorithms proposed in this chapter. The optimum tolerance ϵ in (5.3) for noisy speech in this evaluation, ϵ_o was determined experimentally with babble noise across all codecs, noise levels, and speech to be used in the tests. This was achieved by first generating the ROC curve based on the FPR and FNR values for the algorithm for different values of ϵ, where $\mathcal{FPR}$ is the ratio of samples incorrectly identified as clipped to the total number of unclipped samples and FNR is the ratio of samples incorrectly identified as unclipped to the total number of clipped samples. The optimum value for ϵ is given by the intersection of the ROC curve with the EER. Figure 5.5 indicates that $\epsilon_o = 0.72$.

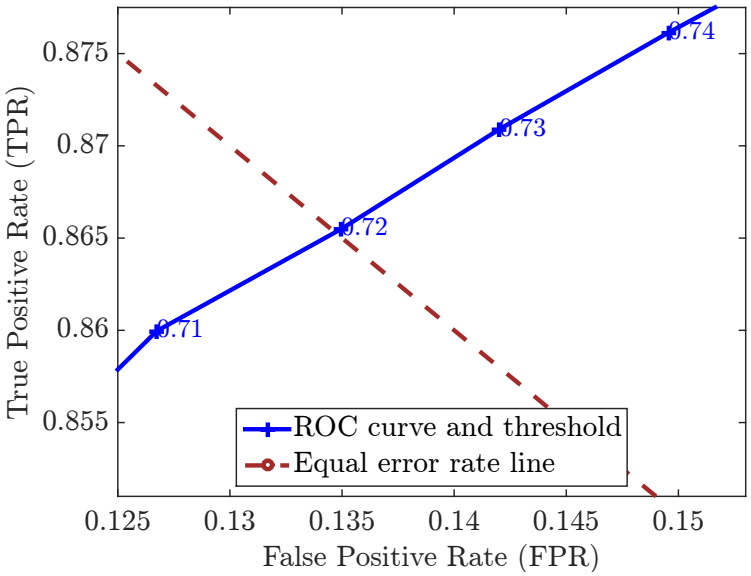

Figure 5.5: ROC curve for identification of optimum threshold for the baseline method in noisy clipped decoded and unencoded speech.

Computational complexity of the algorithms was compared using the RTF for each algorithm calculated as follows: the mean elapsed processing time using the Matlab *tic* and *toc* functions was divided by the mean speech signal duration over all tests to give RTF. Tests implemented in Matlab were run consecutively on an Intel Xeon E5-2643 3.3 GHz processor.

Algorithm	Null codec		GSM		AMR		Overall	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Optimized Baseline	0.843	0.369	0.905	0.460	0.905	0.460	0.885	0.430
ILAH	0.992	0.923	0.944	0.313	0.944	0.313	0.961	0.578
LSR	0.962	0.680	0.959	0.375	0.959	0.375	0.963	0.554
LILAH	0.999	0.992	0.960	0.368	0.960	0.368	0.975	0.667
CRACD	0.997	0.973	0.942	0.521	0.942	0.521	0.957	0.644
ILAH2	0.993	0.930	0.947	0.318	0.947	0.318	0.962	0.588

Table 5.2: Overall ACC and F1 by algorithm and codec for noisy clipped speech in babble noise at 20 dB SNR

5.8 Results and discussion

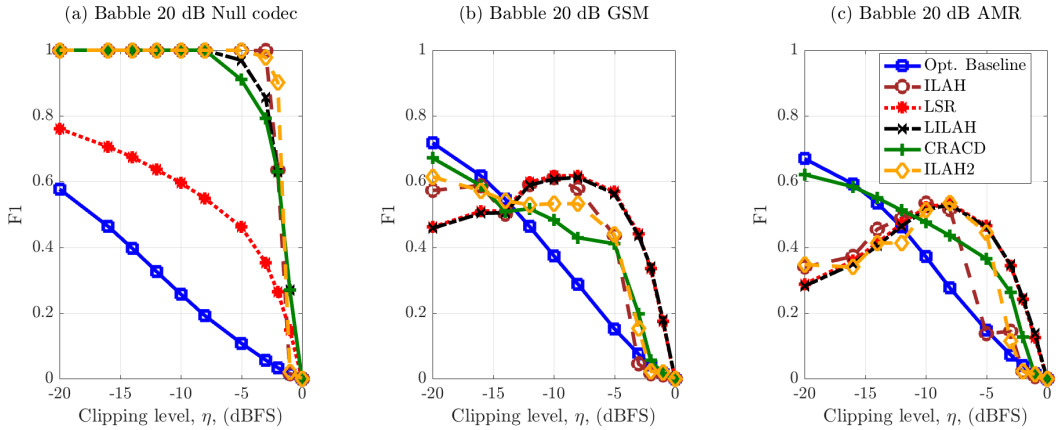

Figure 5.6: Mean F1 with babble noise at 20 dB SNR for (a) the null codec, (b) GSM 06.10 codec, and (c) AMR codec at $\eta = 0$ to -20 dBFS

Figures 5.6, 5.7, 5.9, and 5.10 show the mean F1, ACC, and ACC variance results, Figs. 5.8 and 5.11 show the mean FPRfor each algorithm and clipping level for the null codec, GSM 06.10 codec, and AMR codec for SNRs of 20 dB, and 0 dB respectively, computed from (5.53). Overall mean ACC and F1 results for each algorithm are shown in Table 5.2. The proposed algorithms are compared with the baseline described in Section 5.7.2 optimized for noisy conditions ($\epsilon_o = 0.72$). Results for G.711 are not significantly different to the the null codec case and the are therefore not shown.

As SNR decreases, mean ACC and F1 with the null codec, GSM 06.10 codec, and AMR codec degrade slowly. The algorithm with the highest overall mean ACC and F1 result for unencoded and decoded speech is LILAH. ILAH2, however, without training achieves an

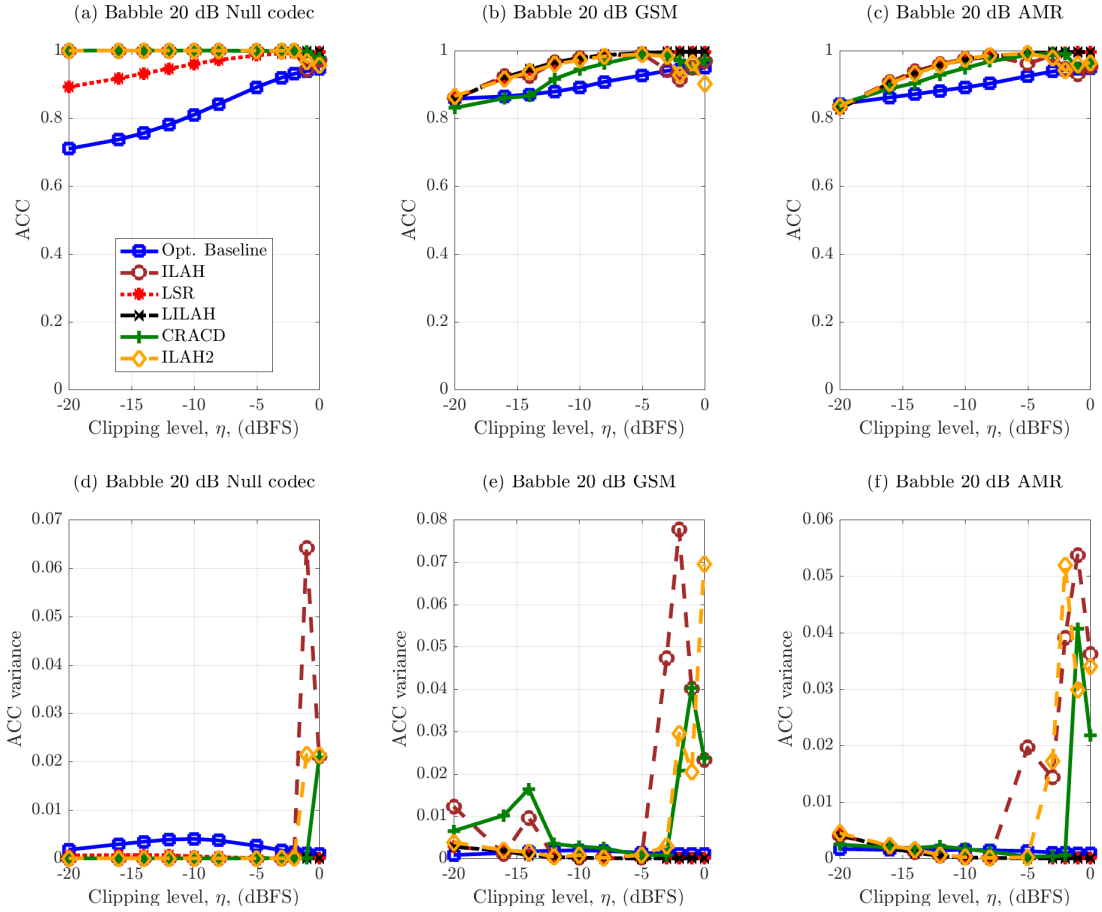

Figure 5.7: Mean ACC and ACC variance with babble noise at 20 dB SNR for (a) and (d) the null codec, (b) and (d) GSM 06.10 codec, and (c) and (f) AMR codec at $\eta = 0$ to -20 dBFS

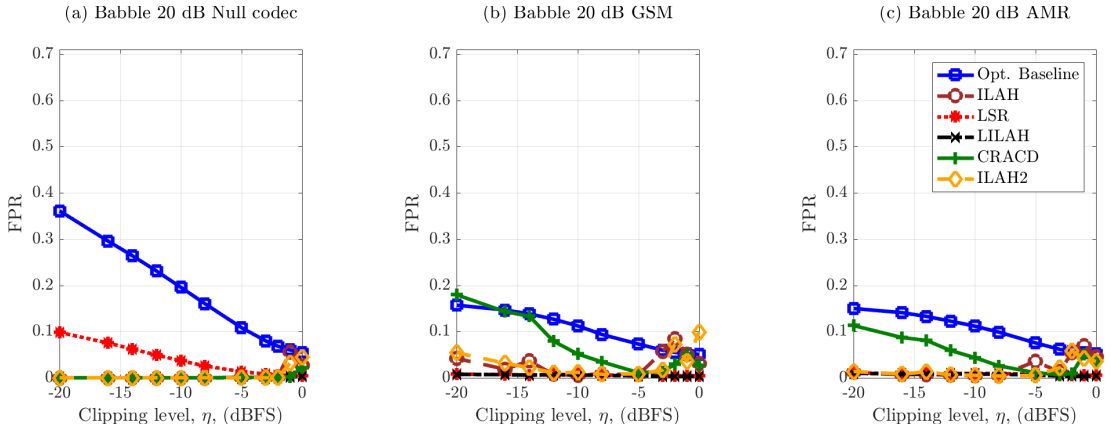

Figure 5.8: FPR with babble noise at 20 dB SNR for (a) the null codec, (b) GSM 06.10 codec, and (c) AMR codec at $\eta = 0$ to -20 dBFS in noise at 20 dB SNR.

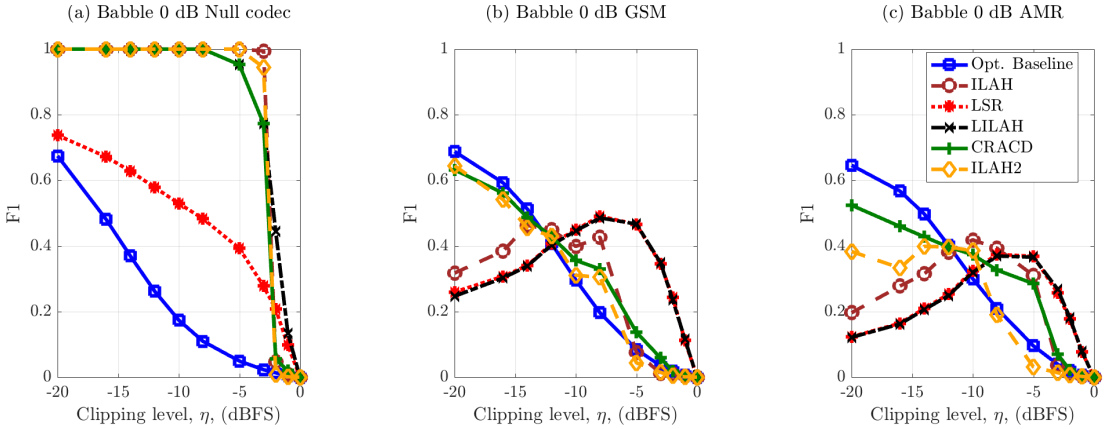

Figure 5.9: Mean F1 with babble noise at 0 dB SNR for (a) the null codec, (b) GSM 06.10 codec and c) AMR codecs at $\eta = 0$ to -20 dBFS.

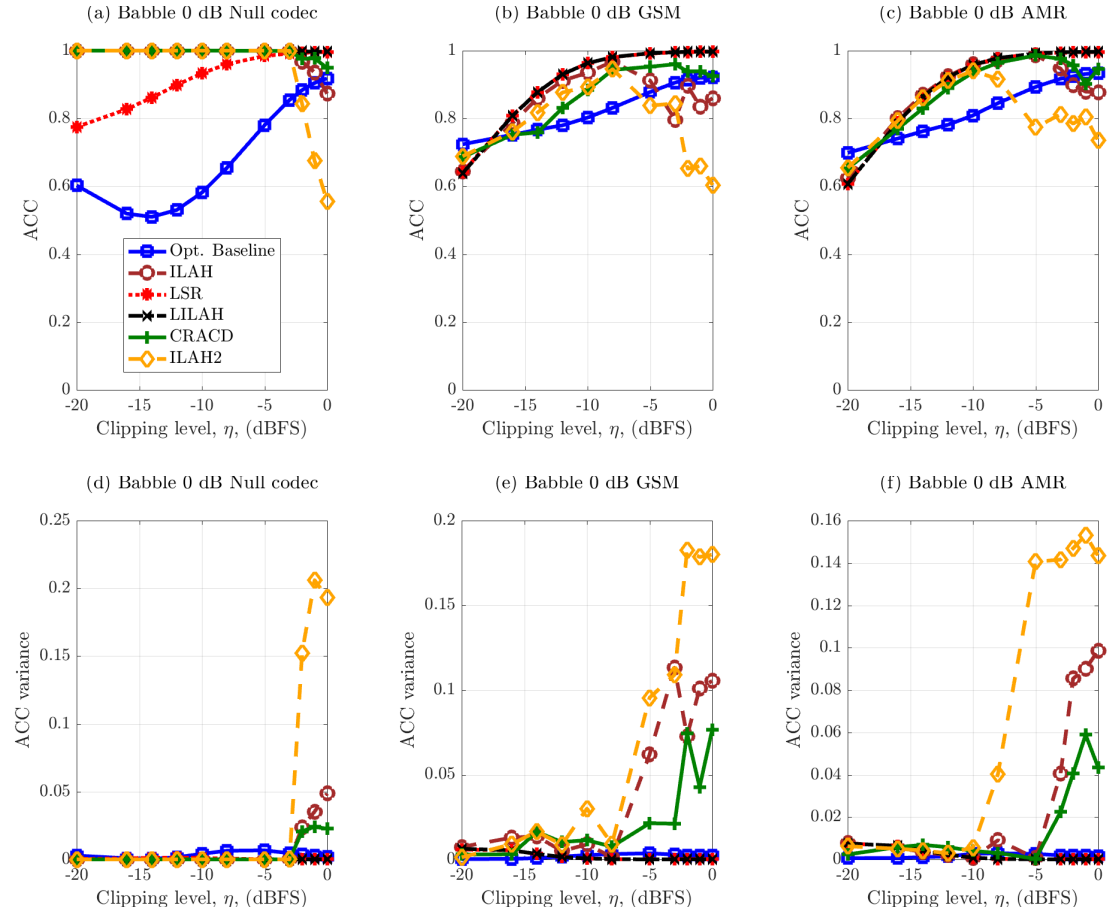

Figure 5.10: Mean ACC and ACC variance with babble noise at 0 dB SNR for (a) and (d) the null codec, (b) and (e) GSM 06.10 codec, and (c) and (f) AMR codec at $\eta = 0$ to -20 dBFS.

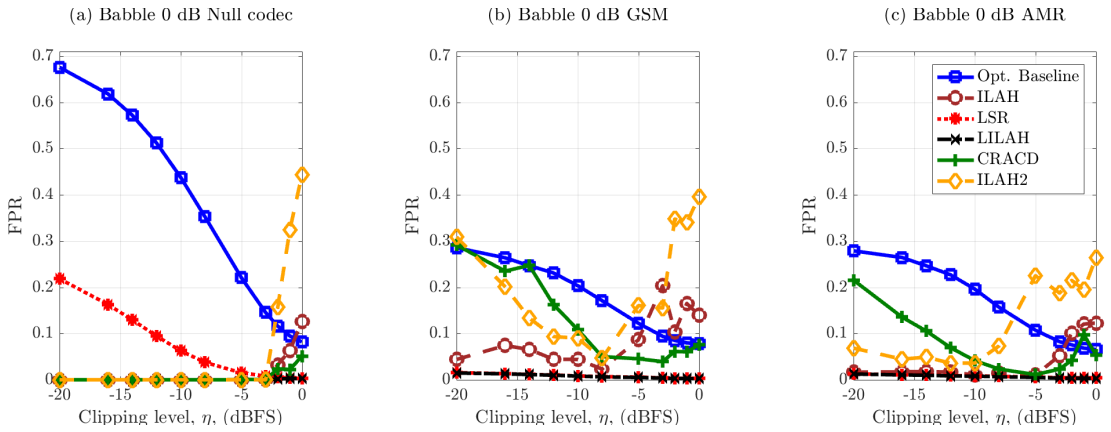

Figure 5.11: FPR with babble noise at 0 dB SNR for (a) the null codec, (b) GSM 06.10 codec and (c) AMR codec at $\eta = 0$ to -20 dBFS in noise at 0 dB SNR.

overall ACC and F1 result between that of LSR and LILAH at 20 dB SNR.

The ACC variance for the Optimized Baseline is much greater than any of the other algorithms with the null codec, and with AMR it is higher for $\eta > -10$ dBFS. This shows that the proposed algorithms adapt to the characteristics of each utterance whilst the baseline does not, as expected. Whilst the Optimized Baseline F1 is comparable to, or better than, the proposed algorithms in perceptually decoded speech at $\eta < -10$ dBFS, in all cases the Optimized Baseline has a higher FPR. The Optimized Baseline is thus erroneously identifying many more unclipped samples as clipped than the proposed algorithms as shown in Figs. 5.8 and 5.11.

The ILAH, LILAH and ILAH2 algorithms provide an estimate of the clipping level and peak absolute unclipped signal level as discussed in Section 5.2 which are useful in clipping restoration algorithms such as [18]. Original signal level estimation results for unencoded and AMR coded clipped signals are shown in Fig. 5.12. Results for GSM 06.10 codec do not differ significantly. Mean estimation error is approximately 10% of signal level with the null codec and with the AMR codec. Larger variances occur with higher η where the long tail of the speech amplitude histogram is close to zero, and small variations in histogram frequency result in larger variations in the iterated logarithm domain.

Results for RTF normalized to the Optimized Baseline of 6.3×10^{-4} are shown in Table 5.3. ILAH and ILAH2 approach this low complexity. The RTF of ILAH2 is higher than ILAH because it performs more least squares regression to determine the best pair of gradients, rather than using predetermined start and end points for the gradients. The

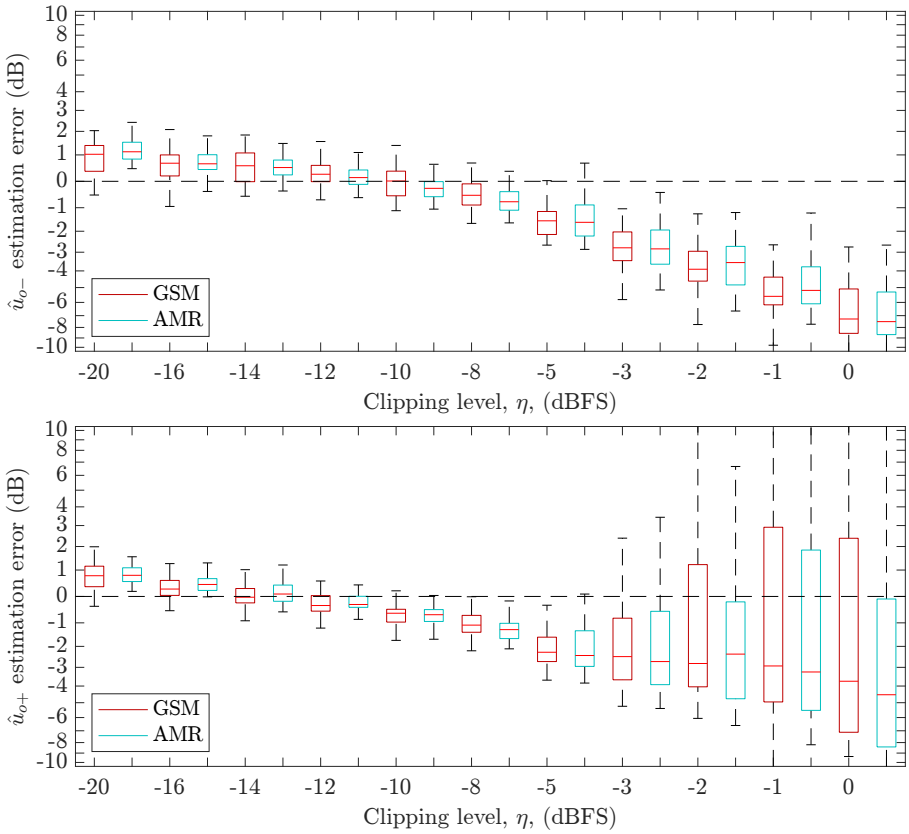

Figure 5.12: ILAH unclipped positive and negative peak absolute signal level, $\hat{u}_{o+}$, and $\hat{u}_{o+}$ respectively estimation accuracy for the null codec and AMR by clipping level in babble noise at 20 dB SNR.

RTF of LSR and LILAH are significantly higher than that of ILAH and ILAH2, although they can still be performed in real-time even in a Matlab implementation. The parameter ϵ in (5.3) does not significantly affect RTF of the Optimized Baseline. Clipping level, η, does not affect RTF for the Optimized Baseline and ILAH.

Optimized Baseline	ILAH	LSR	LILAH	ILAH2
1	2.53	20.4	22	6.09

Table 5.3: Mean RTF by algorithm normalised to the Optimized Baseline for noisy clipped speech

5.9 Conclusion

Determining which samples are clipped in perceptually decoded speech is difficult because the encoding and decoding process modifies both time- and frequency-domain features of clipping. The results above have demonstrated that by using a combination of time and frequency domain features, the LILAH clipping detection algorithm is able to identify clipped samples over the test range of η in AMR decoded speech in babble noise at 20 dB SNR with an overall F1 of 0.758 compared to the Optimized Baseline F1 of 0.406. However, in order to achieve this level of performance, many training steps were required resulting in a large number of implementation dependent parameter. The ILAH2 algorithm however, provides a comparable level of performance to LILAH with training required only to provide an improved estimate of the original unclipped signal level.

ILAH2 outperforms the Optimized Baseline in both unencoded and decoded speech in noisy conditions across a range of clipping levels, noises, noise levels and codecs in the majority of conditions down to 0 dB SNR. Whilst no steps were taken in the design of the presented algorithms to mitigate against noise, the experiments have shown that they are inherently robust to noise, with ILAH2 showing marginally lower robustness to noise than ILAH, LSR and LILAH.

The ILAH2, algorithm performs best with clipping levels between -14 and -3 dBFS which is an important range in typical telecommunications applications, where clipping levels lower than -14 dBFS may be unlikely. Under these circumstances, the baseline crudely labels all samples of significant amplitude as clipped, and is therefore often correct by default. The low computational complexity of ILAH2 makes it suitable for real-time applications.

As bi-products of the detection process, it has been shown that it is possible to estimate the original peak amplitude of the speech signal which may be useful for recovering the clipped speech. It has also been shown that whether or not a perceptual codec has been applied to the speech signal can be reliably estimated.

Chapter 6

The ACE Challenge - Corpus description and performance evaluation

6.1 Introduction

THE acoustic parameters Reverberation Time (T_{60}), and Direct-to-Reverberant Ratio (DRR), together can provide important information about the acoustic environment, and the quality and intelligibility of speech observed in that environment. The performance of speech processing algorithms such as dereverberation [16, 17, 61, 163], and speech recognition [60], can be improved through the use of these parameters.

As was discussed in Chapters 3 and 4, the T_{60} is defined as the time taken for a sound in the diffuse sound field to decay by 60 dB after an abrupt cessation of the source, whilst the DRR is the energy ratio of the sound arriving at the observation point directly from the source to the sound arriving at the observation point after being reflected from one or more surfaces.

The T_{60} and DRR can be estimated from a measured Acoustic Impulse Response (AIR) using existing methods [62, 164]. However, in many practical situations, the AIR is not available and so the T_{60} and DRR must be estimated non-intrusively from the observed reverberant speech. In this context, non-intrusively means that no prior information about

the speech or the AIR is available.

Increasingly, additive noise is a problem for speech processing applications [22]. Gaubitch *et al.* [109] showed that three state-of-the-art T_{60} estimators [77, 86, 87] were significantly biased in the presence of additive noise. Since this review, further developments have been made in non-intrusive T_{60} estimation in noise such as [1, 89, 90, 165, 166]. Further, there have been advances in DRR estimation [5, 118, 119, 167] which have not yet been the subject of a comparative evaluation. In addition, multi-channel devices are now commonplace with typical mobile telephone voice processing components supporting up to the three microphones [168]. Therefore spatial information is available in practical applications, which is useful in determining DRR [5, 167]. There is therefore a need to assess the state-of-the-art in non-intrusive T_{60} and DRR estimation in single and multi-channel scenarios with additive noise.

6.1.1 Accurately simulating additive noise

In order to develop and test acoustic parameter estimation algorithms robust to noise such as [1], it is necessary to have available test signals from either real or simulated noisy reverberant environments. It is usually impractical to record speech in real acoustic environments since each speech utterance needs to be recorded in each acoustic environment which may involve many combinations of utterances, rooms, microphone positions, and noise levels. It is then restrictive in terms of creating new combinations of reverberant speech. For practical purposes, an assumption is therefore made that the AIR represents a Linear Time Invariant (LTI) system, which is a useful approximation for situations where the room conditions are static over the period of time necessary to make AIR measurements. Under this assumption it is possible to construct representative noisy reverberant speech by convolving anechoic speech recordings with measured AIRs, and mixing with noise.

An existing technique for introducing noise is to use pre-recorded monaural noise recordings for the noise source such as [20], and adding it to a speech signal that has been convolved with an AIR. How noise arrives at the microphone however is dependent on the characteristics of the room, and there will be a different AIR for every point from which noise emanates. The noise types of interest in testing of T_{60} and DRR estimation algorithms are numerous. In this study, we have considered in particular babble noise, ambient noise

and fan noise. Babble noise is a common test for speech processing algorithms, representing a realistic scenario of speech in a crowded room, and is applicable to many practical applications including mobile communications and hearing aids. Ambient noise represents the background encountered in everyday situations usually in buildings. Fan noise represents situations where there is close proximity to a computer fan or air conditioning system, for example.

Accurately simulating different types of noise where the noise emanates from many or all parts of a room requires the capture or simulation of a very large number of AIRs. Simulating realistic room noise is generally only practical when adding perfectly diffuse noise sources. It is impractical, therefore, for adding noise emanating from many sources in the same room such as babble noise and non-stationary ambient noise since many individual measurements would need to be made for each source.

Of the corpora used in recent T_{60} and DRR estimation research [98, 101, 102, 108], there are no associated noise recordings. In a related study, the Reverb Challenge, the SimData corpus [169] incorporates ambient noise recorded under the same conditions as the AIRs. A further recent corpus, Distant-speech Interaction for Robust Home Applications (DIRHA) [170] incorporates recordings of domestic noises from rooms within a single apartment. Whilst ambient noise is provided in two of the corpora above, neither of these studies provides babble noise and fan noise recorded under the same conditions as the AIR measurements.

6.1.2 ACE Challenge and corpus

In order to stimulate research and to determine the state-of-the-art in non-intrusive acoustic parameter estimation in realistic noise, inspired by [109], the Acoustic Characterization of Environments (ACE) Challenge was devised involving the collection of a new corpus of noisy reverberant speech. The ACE corpus comprises spontaneous anechoic speech, measured multi-channel AIRs, and multi-channel ambient, fan and live babble noise recorded under the same conditions as the AIRs. The ACE Challenge is one of several research challenges coordinated by the IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee.

The contribution of this chapter is to describe the ACE Corpus and how it was

constructed, describe the process of the challenge, describe the use of the corpus in the ACE Challenge to perform comparative testing of existing and novel algorithms on the same data including the algorithms proposed in Chapters 3 and 4, and to provide a comparative evaluation and analysis of the results.

The remainder of the chapter is organised as follows: In Section 6.2, ACE Corpus is described, along with the methods for determining the ground truth T_{60} and DRR. In Section 6.3, the ACE Challenge is described. In Section 6.4, a summary of the results of the ACE Challenge are provided including the results of the algorithms proposed in Chapters 3 and 4, and in Section 6.5, conclusions are drawn.

6.2 ACE Corpus

The ACE Corpus[1] is a new database of anechoic speech, multi-channel AIRs, and multi-channel noise for providing realistic multi-channel noisy reverberant speech utterances for the development and evaluation of speech processing algorithms. Included are AIRs for seven different rooms with five different microphone array configurations in two different microphone positions per room in order to provide a wide range of T_{60}s and DRRs. Importantly, it also includes babble, ambient and fan noise recorded under same conditions as the measured AIRs. The T_{60} and DRR were obtained from the measured AIRs for each room, microphone position, and microphone, in both Fullband (FB) and ISO preferred frequency bands [59], and these ground-truth parameter values are provided with the corpus. In addition, a set of anechoic recordings of spontaneous speech is provided. Realistic noisy reverberant speech signals can be constructed by combining the various elements of anechoic speech, AIRs and associated noise using software supplied with the corpus. These signals can span a wide range of T_{60}, DRR and Signal-to-Noise Ratio (SNR), and, whilst the corpus has been designed specifically to test acoustic parameter estimation algorithms, it is also highly relevant to many other real-world speech processing tasks.

[1] Available `www.ace-challenge.org`

6.2.1 Acoustic model

The acoustic model in the Chapter extends the model introduced in Chapter 3, Section 3.1, to operate with observations from multiple microphones. A noisy reverberant speech signal, $y_m(n)$, with discrete-time index, n, captured in a room by a microphone m, is characterised by an AIR, $\mathbf{h}_m$, of length L_h, of the acoustic channel between the source and microphone, where $\mathbf{h}_m = [h_{m,0}, h_{m,1}, \ldots, h_{m,L_h-1}]^T$. It is assumed that the AIR is time invariant. The noisy reverberant signal incorporating the convolution of the speech with the AIR can be written as

$$y_m(n) = \mathbf{h}_m^T \mathbf{x}(n) + v_m(n), \tag{6.1}$$

where $\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-L_h+1)]^T$ is the speech signal vector, and $v_m(n)$ is additive noise at the m^{th} microphone. To obtain a noisy reverberant speech signal, $y_m(n)$, at the test SNR, ξ, a clean speech signal, $x(n)$, is convolved with the respective AIR, $\mathbf{h}_m$, as in (6.1), and then mixed with noise as

$$y_m(n) \quad = \quad \mathbf{h}_m^T \mathbf{x}(n) + \frac{v_m(n)\sqrt{\xi_s}}{\sqrt{\xi}} \tag{6.2}$$

where ξ_s is power ratio of the unprocessed speech and noise signals given by

$$\xi_s = \frac{A_x}{E\{|v_m(n)|^2\}}, \tag{6.3}$$

and where A_x is the active speech level of the reverberant speech utterance without noise estimated using ITU-T P.56 Method B [25], and $E\{\cdot\}$ is the expectation operator.

6.2.2 Anechoic speech recordings

Two sets of speech signals were recorded. In Set 1, four male talkers were recorded and in Set 2, five female and five male talkers were recorded. The utterances comprise the subject speaking their favourite colour, the town where they live, a description of where they live, a description of how they travel to work, and counting from zero to nine. The utterances are in different dialects of international English with a mix of native and non-native English speakers. Recordings were performed in the anechoic chamber at TU Delft in a single sitting per talker per set. A B&K 4190-L-001 measurement microphone [171] connected to

a B&K NEXUS 2690 conditioning amplifier was used, and digitized using an RME FireFace 800 audio interface, using a sample rate of 48 kHz and 24-bit precision. The speech signals were then normalised to give approximately equal loudness measured in Loudness Units Full-Scale (LUFS), and then manually segmented into utterances which vary in length according to utterance type and speaker. Following the EBU recommendation R 128 [172], given a target loudness of −23 LUFS, the normalization ensured all speech signals were within ±1 LUFS of the target.

6.2.3 Rooms

Table 6.1 lists the dimensions for the seven different offices, meeting and teaching rooms within Imperial College London that were used to produce the corpus. Also included for each room is the volume, estimated mean T_{60} across all microphones and AIRs, and the range of estimated DRR for the two microphone positions, Position 1 and Position 2. The room characteristics are as follows:

Office 1: a carpeted office containing a table, desk and four chairs;

Office 2: a carpeted office containing a table, desk, 6 chairs and a bookcase;

Meeting Room 1: a carpeted meeting room containing a meeting table and 14 chairs;

Meeting Room 2: a carpeted meeting room containing approximately 30 chairs and 6 tables;

Lecture Room 1: a hard-floored lecture room containing approximately 20 tables and 60 chairs;

Lecture Room 2: a hard floored lecture room containing approximately 35 tables and 100 chairs;

Building Lobby: a large irregular-shaped hard-floored room with coupled spaces including a café, stairwell and staircase (shown in Fig. 6.1). Measurements in Table 6.1 correspond to the corner area where the recordings were made whereas the total volume of the lobby is many times larger. There was a high level of both stationary and non-stationary naturally occurring ambient noise including electrically operated doors, lifts with announcements, and building users walking by the recording environment.

Name	L (m)	W (m)	H (m)	Vol. (m^3)	T_{60} (s)	Pos. 1 DRR min. (dB)	Pos. 1 DRR max. (dB)	Pos. 2 DRR min. (dB)	Pos. 2 DRR max. (dB)
Office 1	4.83	3.32	2.95	47.3	0.332	−2.73	13.0	−0.551	6.56
Office 2	5.10	3.22	2.94	48.3	0.390	−0.444	13.0	−2.28	9.48
Meeting Room 1	6.61	4.73	2.95	92.2	0.437	−1.98	10.8	−3.10	7.57
Meeting Room 2	10.3	9.17	2.63	249	0.371	−2.57	11.2	1.08	12.5
Lecture Room 1	6.93	9.73	3.00	202	0.638	−0.825	14.6	0.872	7.90
Lecture Room 2	13.4	9.29	2.94	365	1.22	−0.370	12.8	−3.75	6.45
Building Lobby	5.13	4.47	3.18	72.9	0.646	−0.936	13.0	−2.50	8.09

Table 6.1: Room dimensions (approx.), mean FB T_{60}, and mean FB DRR across all microphone positions, configurations, and channels

6.2.4 Microphone positions, configurations, source and seating positions

The source position and the seating position of the occupants remained the same for all recordings within each room, whilst there were two separate positions per room of all the microphones, Position 1 and Position 2. Table 6.2 shows the microphone configurations used for all AIR and noise recordings in the corpus. All recordings were made sample-

Name	Chs.	Type	Alignment	Spacing	Elements
Chromebook	2	Chromebook Pixel	Horizontal	Laptop screen bezel, 62mm	MEMS
Mobile	3	Perspex former	Right-angled triangle	Base:45mm, side:100mm	DPA4060 [173] Omni
Crucif	5	Perspex former	4-arm with central mic.	250mm centre-to-arm	DPA4060 [173] Omni
Lin8Ch	8	Perspex former	Linear	60mm	DPA4060 [173] Omni
EM32	32	Eigenmic [174]	Spherical	Dist. on 84 mm rigid sphere baffle	14mm electret pressure

Table 6.2: ACE Challenge Microphone configurations

synchronously across all channels using a sample rate of 48 kHz and 24-bit precision.

The Mobile, Crucif and Lin8Ch arrays were recorded using two RME OctaMic preamps with their balanced outputs connected to the balanced inputs of two clock-synchronized RME FireFace 800s connected to a MacBook Pro. The recording software employed was Audacity [175]. Eigenmike recordings were made using the Eigenmike

Microphone Interface Box (EMIB) clock-synchronized to the FireFace 800s connected to a second MacBook Pro. The Chromebook, recordings were made using *arecord* in little endian format so as to directly record the signals from the two integrated microphone, but not however clock-synchronised to the other audio interfaces.

6.2.5 Room recording procedure

The recording procedure in each room involved the following steps:

1. Install recording equipment positioning the microphones in Position 1, and document the room dimensions and the positions of all microphones, sources and seats;
2. Make empty-room AIR measurements and noise recordings; Empty-room measurements were for verification purposes and do not form part of the published corpus since the set is not complete, although they may be used in future experiments;
3. Participating subjects take their seating positions;
4. Make occupied AIR measurements and noise recordings;
5. Move microphones to Position 2 and document their positions;
6. Make occupied AIR measurements and noise recordings;
7. Participants leave the room;
8. Make unoccupied AIR measurements and noise recordings in the second microphone position;
9. Uninstall recording equipment.

Figure 6.1 shows the recording equipment in place in the Building Lobby ready to commence recording.

6.2.6 Noise recordings

Three different noise types (ambient, fan and babble) were recorded in each room for each microphone position. To generate the babble noise, four to seven people sitting in

Figure 6.1: Recording session in the Building Lobby before the occupants arrive with microphones in Position 1

the recording environment were asked to speak continuously for the duration of the noise recording. Talkers were provided with a list of phrases from TIMIT [19], or could bring their own material. In a few cases talkers read from scientific papers. To maintain as constant an acoustic environment as possible, talkers remained in their seated positions for the duration of the recording session of all noises and all AIR measurements. The ambient noise was recorded with occupants present but remaining silent. One or two fans in the room were used to create the fan noise, taking care to avoid creating wind noise.

For all of the rooms and microphone positions except Office 2, AIRs for the rooms without participants were captured. The effect of removing the talkers is discussed in Section 6.3. For both the fan and babble noises, there is also ambient noise in the background since this could not be avoided. To provide similar noise across all microphones, noise recordings from the Chromebook, FireFace 800s and EMIB were aligned so that channel 1 of the first FireFace 800, channel 1 of the Chromebook and channel 1 of the Eigenmike start at the same time, using *sigalign.m* [37]. The signals were then trimmed so that the noise recordings start and finish at the same point in time across all interfaces.

6.2.7 AIR measurement

The exponential frequency sweep method of Farina *et al.* [65,66] was used to measure the AIRs. In this method, a test signal, $x(n)$, is played through an amplifier and loudspeaker placed in the specified source position. The signal, $y(n)$, which includes the effect of the room, recorded using a microphone at the specified sensor position, can be expressed as

$$y(n) = \mathbf{h}^T \mathbf{x}(n), \tag{6.4}$$

where $\mathbf{h}$ is the unknown AIR, $\mathbf{x}(n)$ is defined as in (6.1), and the excitation signal, $x(n)$, is given by

$$x(n) = \sin\left[\frac{2\pi f_1 N}{f_s \log\left(\frac{f_2}{f_1}\right)}\left(\exp^{\left\{\frac{n}{N}\log\left(\frac{f_2}{f_1}\right)\right\}} - 1\right)\right]. \tag{6.5}$$

Equation (6.5) is an exponential sweep from frequency f_1, to frequency f_2, with a duration of N samples or N/f_s seconds, where f_s is the sample rate in Hertz. An inverse filter, g, can be found satisfying

$$\delta(n) = \mathbf{g}^T \mathbf{x}(n), \tag{6.6}$$

where $\delta(n)$ is the delayed Dirac delta function, $\mathbf{g} = [g_0, g_1, \ldots, g_{L_g-1}]^T$, and L_g is the length of the inverse filter. In an anechoic environment, and assuming no colouration from any of the transducers

$$\delta(n) = \mathbf{g}^T \mathbf{y}(n), \tag{6.7}$$

where $\mathbf{y}(n) = [y, y(n-1), \ldots, y(n-L_g+1)]^T$. However, in a real acoustic environment, $y(n)$ will be convolved with the AIR, thus the AIR can be found as

$$\mathbf{h} = (\mathbf{g}^T \mathbf{y}(n))^T. \tag{6.8}$$

The exponential sweep in (6.5) ensures that harmonic distortions appear in $\mathbf{h}$ at precise time lags before the AIR, and produces a better SNR at low frequencies [65]. The inverse filter is generated by time-reversing the input signal $x(n)$. The convolution was performed efficiently in the frequency domain using the Fast Fourier Transform (FFT). The excitation signal was played through a Fostex 6301B Personal Monitor loudspeaker and recorded using all microphone configurations simultaneously. The tail of each AIR was faded down to zero over a period of 208 ms (10 000 samples) once the level fell below −70 dB to remove

noise and artefacts. The AIRs were analysed for FB and Subband (SB) T_{60} and DRR as will be described in Section 6.2.8, and 6.2.9 respectively. Table 6.1 shows for each room the mean values for T_{60} over all microphones, channels and positions for each room, and minimum and maximum DRR.

6.2.8 Obtaining T_{60} estimates from the AIR

The method of Karjalainen *et al.* [164] was used to determine the ground truth T_{60} for each channel of each impulse response measurement in FB and in frequency bands using the ISO preferred centre frequencies [59]. Thus the centre frequency of band 1 is at 25.1189 Hz, band 7 is at 100 Hz, band 17 is at 1000 Hz, band 27 is at 10 kHz, and band 30 is at 19.953 kHz. This method applies a non-linear optimization to a model for the reverberant signal which includes three components: the exponential decay of the reverberant signal as a single mode, the reverberant tail in the diffuse sound field, and the noise floor. This was found to be more reliable under all conditions than either the ISO-3382 T_{30} or T_{20} derived T_{60} estimates [64] which tended to over-estimate the reverberation time in the rooms. The filter bank used for the frequency-dependent T_{60} estimation was a one third-octave 8^{th} order Butterworth design, designed using the Matlab *fdesign.octave* function. Centre frequencies were generated using the *validfrequencies* function.

Figure 6.2 shows T_{60}s calculated from different AIR measurements made within a single room. It can be seen that in the range 1 kHz to 5 kHz, the T_{60} is approximately 200 ms larger for the unoccupied room compared to the occupied room, showing that the participants in the experiment absorbed sound energy and reduced the reverberation in the room. This shows that it was necessary to measure the AIRs and record all noises with the occupants in the room to maintain the same environmental characteristics.

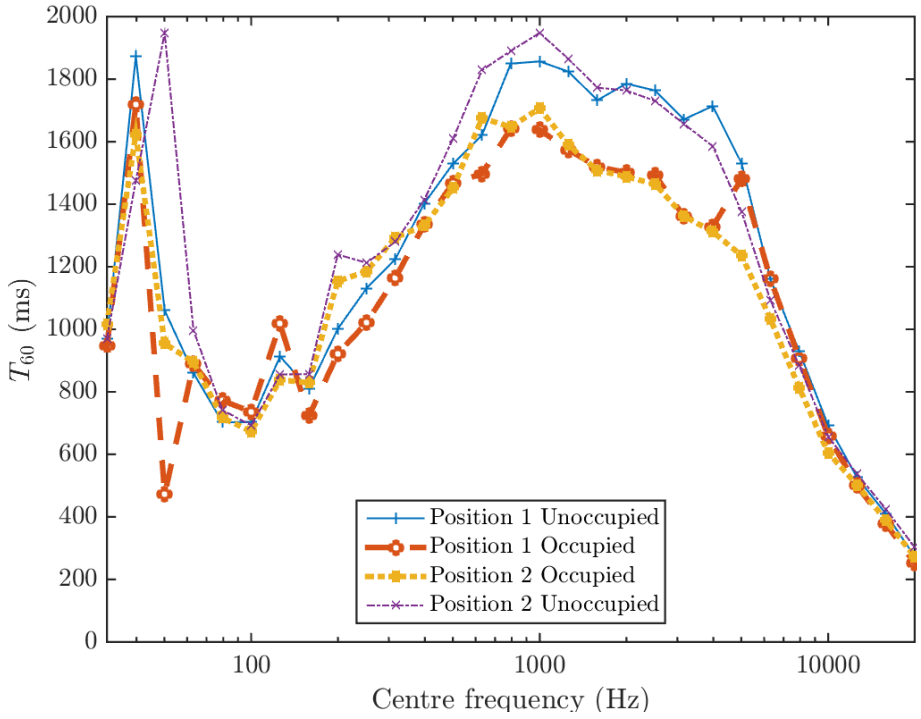

Figure 6.2: Frequency-dependent T_{60} measurements for Lecture Room 2.

6.2.9 Obtaining DRR estimates from the AIR

The DRR at the m^{th} microphone in each microphone configuration was calculated using the method of [67] as

$$\text{DRR}_m = 10\log_{10}\left(\frac{\sum_{n=n_d-n_0}^{n_d+n_0} h_{m,n}^2}{\sum_{n=0}^{n_d-n_0} h_{m,n}^2 + \sum_{n=n_d+n_0}^{n=\infty} h_{m,n}^2}\right), \tag{6.9}$$

where the direct path signal arrives at sample n_d, and $n_0 = 120$ samples at a sample rate of 48 kHz, or 2.5 ms, which represents an additional path difference of 0.85 m at 340 m s^{-1}. In order to identify the sample pertaining to the direct path, n_d, an equalization filter was applied to the AIR which compensated for the frequency response of the source loudspeaker and produced coefficients in the AIR representing distinct reflections. The location of the direct path, n_d, was then determined by finding the maximum value of the AIR after equalization. The AIR without equalization was then used in (6.9). Frequency-dependent DRR estimates were made using the same method as for the frequency-dependent T_{60} estimates described in Section 6.2.8.

6.3 The ACE Challenge comparative evaluation

6.3.1 ACE Challenge overview

The ACE Challenge comparative evaluation was devised to determine the state-of-the-art in non-intrusive acoustic parameter estimation - specifically T_{60} and DRR - and to stimulate research in this field. Central to the challenge was the noisy reverberant speech of the ACE Corpus described in Section 6.2. The following non-intrusive estimation tasks were set:

1. Single microphone FB T_{60} and DRR estimation;
2. Multi-microphone FB T_{60} and DRR estimation;
3. Single microphone T_{60} and DRR estimation in one-third-octave SBs;
4. Multi-microphone T_{60} and DRR estimation in one-third-octave SBs.

Participants were expected to use the Development (Dev) data-set to tune their algorithms for T_{60} and DRR estimation, and then test their algorithms on the fully blind Evaluation (Eval) dataset using the provided software tools. The organizers then decoded and analyzed the submitted results and returned them to the participants.

The ACE Challenge attracted participation from nine research institutions from eight different countries around the world. Participants and their algorithms are listed in Table 6.3 in order of appearance of their algorithms in the results tables. The T_{60} estimation algorithms based on that proposed in Chapter 3 are identified with the letters F and G, and the DRR estimation algorithms based on that proposed in Chapter 4 are identified with the letters j, k, l, m, and n.

6.3.2 ACE Challenge datasets

Dev and Eval datasets of noisy reverberant speech files were constructed from the ACE Corpus described in Section 6.2. The purpose of the Dev dataset was to allow participants of the ACE Challenge to review the performance of their algorithms on typical ACE data, and perform any tuning before commencing the challenge. The purpose of the Eval dataset was to provide non-intrusive noisy reverberant speech upon which to base the challenge.

Participant	Algorithms submitted (see results tables)	
	T_{60}	DRR
Federal University of Rio de Janeiro (UFRJ)	A	y
Friedrich-Alexander-Universität (FAU), RWTH Aachen University	B, C, D, E	
Imperial College London	F, G	j, k, l, m, n
Fraunhofer IDMT, Carl von Ossietzky University, Oldenburg	H	i
INRS-EMT, University of Quebec,	I, J, K, L, M, N, O, P, Q	o, p, q, r, s, t, u, v, w
Nuance Communications, Inc.	R, S, T	e, f, g
Microsoft Research	U, V	
University of Auckland, NTT Corporation		Z, a, b, c, d
Australian National University (ANU)		h

Table 6.3: ACE Challenge participants

Each noisy reverberant speech file was constructed from anechoic speech convolved with measured AIRs obtained from a given room, combined with a random selection of the noise (ambient, fan or babble) recorded in the same room conditions, with the same occupants, and with the same source and microphone configuration using the method described in Section 6.2.1.

The Dev dataset comprised noisy reverberant speech files from 2 rooms with 2 microphone positions each, 4 male talkers, 2 utterances each, babble, fan, and ambient noise at 0 dB, 10 dB, and 20 dB SNR for all microphone configurations. There were 288 files per microphone configuration. Ground truth T_{60} and DRR information for the Dev set was provided to participants for every channel of every microphone position in both FB and in ISO SBs using ISO preferred frequencies [59].

The Eval dataset comprised noisy reverberant speech files from 5 rooms different to those provided in the Dev dataset, with 2 microphone positions each, 5 male and 5 female talkers, 5 utterances each, babble, fan and ambient noise in low, −1 dB, medium, 12 dB, and high 18 dB SNR scenarios generated using the same method as for the Dev dataset. There were 4500 files per microphone configuration. The noisy reverberant speech files were numbered in a pseudorandom permutation, which was different for each microphone configuration, to reduce the possibility of training on the Eval dataset. Both Dev and Eval

datasets were downsampled prior to distribution to a sample rate of 16 kHz and converted to 16-bit precision.

In both Dev and Eval datasets, a further single-channel microphone configuration was included. In the Dev dataset channel 1 of the 8-channel linear array was used. For the Eval dataset channel 1 of the 5-channel cruciform was used.

6.3.3 ACE Challenge software

Software was provided to participants for each phase to assist in generating results in the required format. The Dev software allowed participants to test their algorithms on the Dev dataset and analyse the results by SNR, by T_{60}, and by DRR, from a database of ground truth values. The Eval software allowed participants to test their algorithms on the Eval dataset to produce data files suitable for submission to the ACE Challenge. The software also recorded Real-Time Factor (RTF) for each test, the Direction-of-Arrival (DoA) azimuth and elevation if estimated, and the SNR if estimated for subsequent analysis.

6.3.4 ACE Challenge process

The ACE Challenge was conducted according to the following process:

1. Participants request to participate in the challenge;
2. Organizers release datasets and software to participants;
3. Participants tune their algorithms on the Dev dataset;
4. Participants test their algorithms on the Eval dataset and submit results;
5. Organizers return decoded results to participants with ground truth;
6. Participants submit conference paper to ACE Workshop;
7. Participants present paper describing the algorithm used.

In order to stimulate research in this area, participants were encouraged to enter multiple types of algorithms. They could then select the most promising algorithms to report in their paper submissions.

6.3.5 Taxonomy of algorithms submitted

There were three main classes of algorithms submitted to the ACE Challenge:

1. Analytical, with or without bias compensation;
2. Single feature with mapping;
3. Machine learning using multiple features.

Analytical approaches derive the estimate for the acoustic parameter directly from the signal without requiring any prior information. Bias compensation may be performed in order to account for noise or specific aspects of the source material. An example of this is the maximum likelihood method [77] which directly produces the T_{60} estimate.

Single feature with mapping approaches estimate a parameter from a signal that is correlated with the acoustic parameter to be estimated, and then apply a mapping function to give the acoustic parameter estimate. An example of this is the Spectral Decay Distributions (SDD) method which determines Negative-Side Variance (NSV) from STFT bins and then applies a mapping to obtain the T_{60}.

Machine learning approaches typically use many features of the source material to train a neural network which then estimates the acoustic parameter from the features of a test signal. An example of this is the Non-Intrusive Room Acoustics (NIRA) [176] algorithm.

There were no hybrid approaches submitted to the ACE Challenge although several participants applied noise reduction to the source signals before performing parameter estimation.

6.4 Results and discussion

In this Section, a synoptic presentation of the comparative results is given. Unabridged results can be found in the ACE Technical Report [12]. Separate performance comparisons are provided for the 4 tasks given in Section 6.3.1. For the FB tasks the results are presented as box plots where there is a box shown for each algorithm. On each box in the box plot, the central mark is the median, the edges of the box are the 25^{th} and 75^{th}

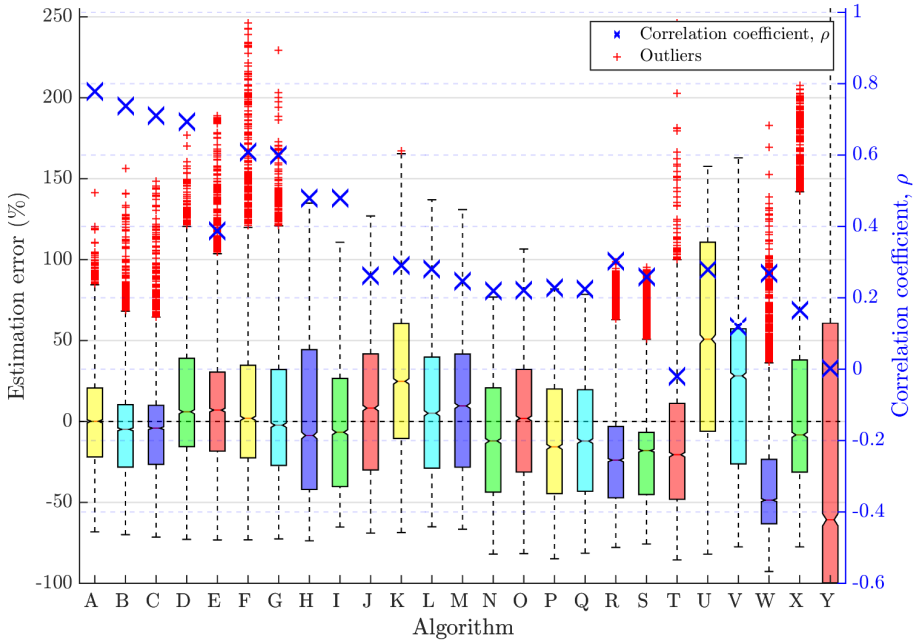

Figure 6.3: FB T_{60} estimation error in all noises for all SNRs

Ref.	Algorithm	Mic. Config.	Bias	MSE	ρ	RTF
A	QA Reverb [177]	Single	-0.068	0.0648	0.778	0.4
B	Octave SB-based FB RTE [178]	Single	-0.104	0.0731	0.738	0.939
C	DCT-based FB RTE [178]	Single	-0.104	0.0766	0.71	1
D	Model-based SB RTE [178]	Single	-0.0363	0.102	0.693	0.451
E	Baseline algorithm for FB RTE [178]	Single	-0.0432	0.11	0.387	0.0424
F	SDDSA-G retrained [8]	Single	0.0167	0.0937	0.608	0.0152
G	SDDSA-G [1]	Single	-0.0423	0.0803	0.6	0.0164
H	Multi-layer perceptron [179]	Single	-0.0967	0.104	0.48	0.0578
I	Per acoust. band SRMR Section 2.5. [180]	Single	-0.114	0.109	0.48	0.578
J	NSRMR Section 2.4. [180, 181]	Single	-0.0646	0.119	0.261	0.571
K	NSRMR Section 2.4. [180, 181]	Chromebook	0.012	0.116	0.291	1.04
L	NSRMR Section 2.4. [180, 181]	Mobile	-0.0504	0.0958	0.281	1.58
M	NSRMR Section 2.4. [180, 181]	Crucif	-0.0516	0.107	0.246	2.62
N	SRMR Section 2.3. [180]	Single	-0.16	0.144	0.22	0.457
O	SRMR Section 2.3. [180]	Chromebook	-0.105	0.132	0.221	0.829
P	SRMR Section 2.3. [180]	Mobile	-0.153	0.12	0.228	1.26
Q	SRMR Section 2.3. [180]	Crucif	-0.153	0.128	0.225	2.09
R	NIRAv3 [176]	Single	-0.192	0.151	0.302	0.899
S	NIRAv1 [176]	Single	-0.184	0.151	0.258	0.899
T	NIRAv2 [176]	Single	-0.179	0.198	-0.0199	0.907
U	Blur kernel [91]	Single	0.173	0.15	0.279	8.46
V	Blur kernel with sliding window [92]	Single	-0.00555	0.139	0.12	0.421
W	Temporal dynamics [86]	Single	-0.304	0.211	0.269	0.362
X	Improved blind RTE [77]	Single	-0.0635	0.165	0.166	0.0259
Y	SDD [87]	Single	0.463	305	0.00158	0.0221

Table 6.4: T_{60} estimation algorithm performance in all noises for all SNRs

percentiles, the whiskers extend to the most extreme data points not considered outliers. Outliers are plotted individually. The algorithms are identified on each box plot by a single character which corresponds to the character in the summary numerical results presented in Tables 6.4 and 6.5. The Pearson correlation coefficient, ρ, between the estimates and the ground truth for each algorithm is plotted as a black cross in the same column as the algorithm. The value is provided on each right hand vertical axis. The results are sorted

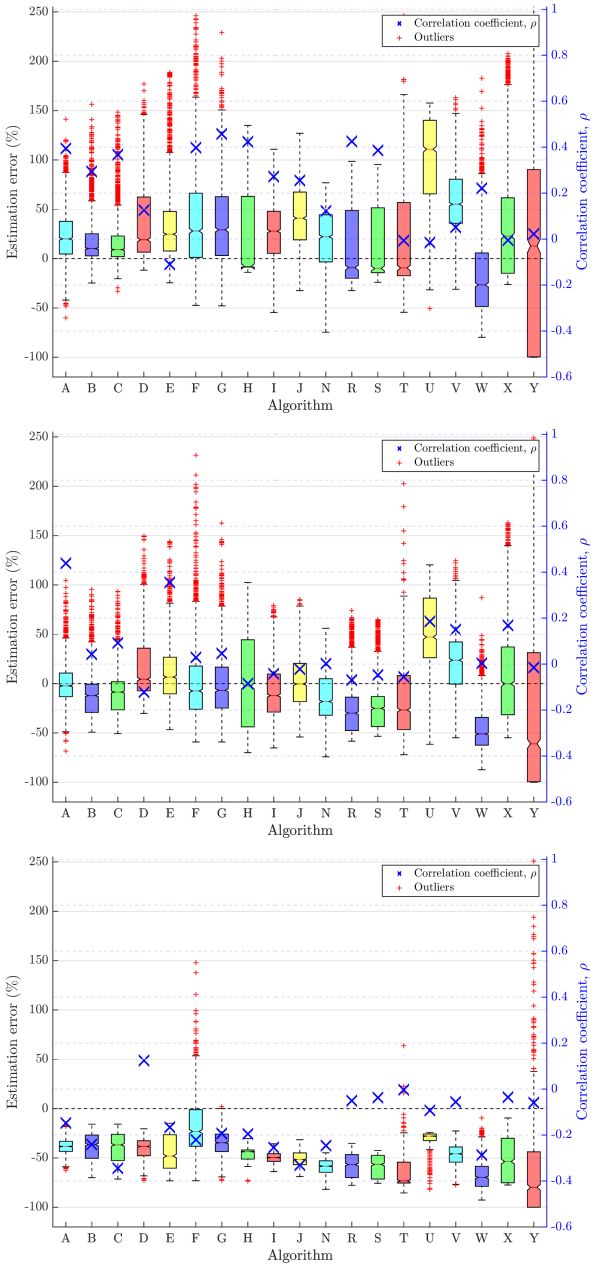

Figure 6.4: Single channel FB T_{60} estimation error in all noises and all SNRs for a) T_{60} <0.43 s b) 0.43 $\leqslant T_{60}$ < 0.75 s and c) $T_{60} \geqslant$ 0.75 s. Observe that $\rho < 0$ for all except algorithm D

by participant according to correlation coefficient across all noises and SNRs in FB. For the T_{60} FB task, the last three algorithms (W, X, Y) are those compared in Gaubitch *et al.* [109], and are included as baselines to enable the progress made in non-intrusive T_{60} estimation since the earlier evaluation to be assessed. Similarly, for the DRR FB task, Jeub *et al.* [118] is included as baseline algorithm since this was a freely available estimator prior

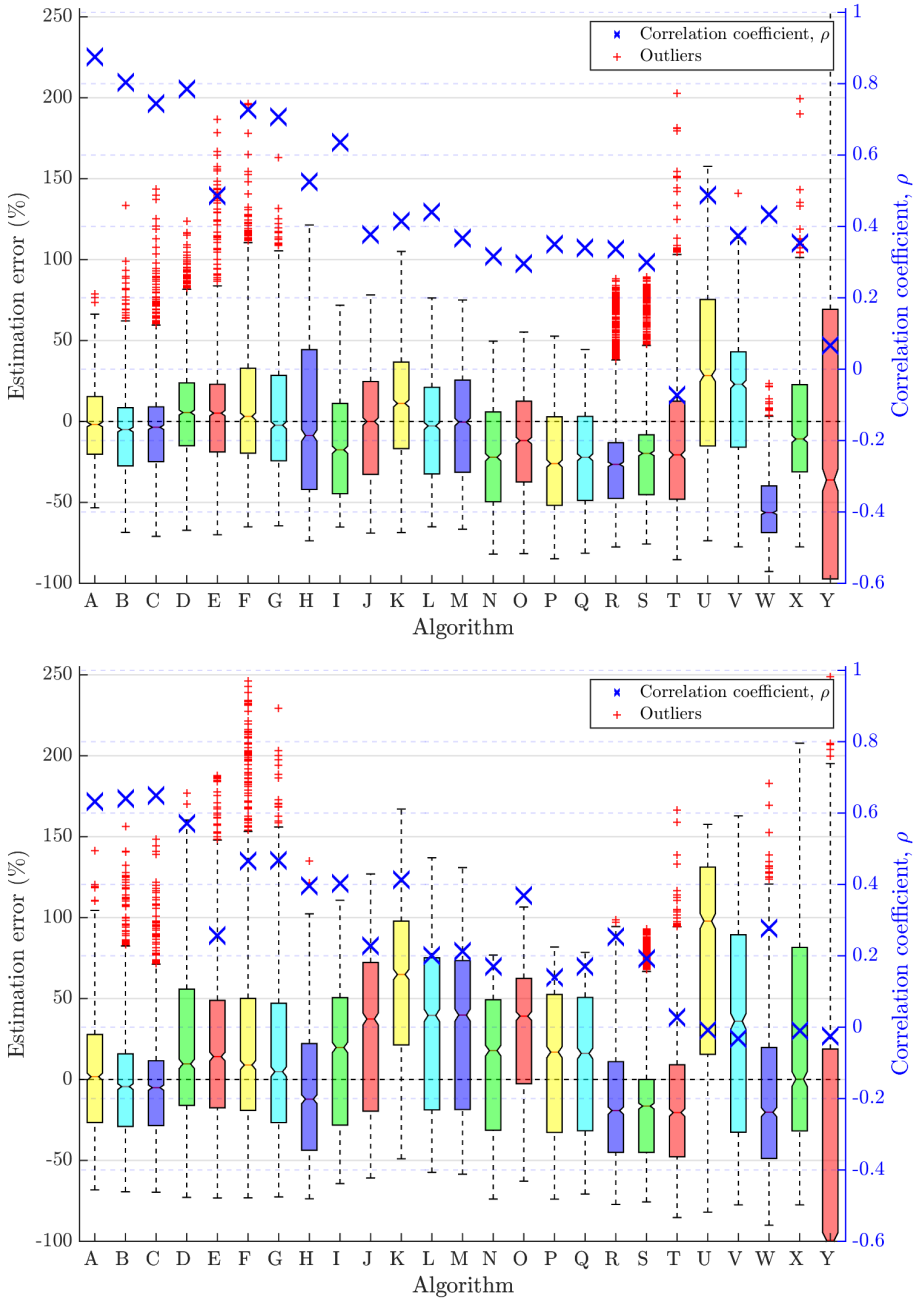

Figure 6.5: FB T_{60} estimation error in all noises at a), 18 dB SNR and b) −1 dB SNR

to the ACE Challenge. The columns in the table are as follows:

1. Ref., the identifier for each algorithm used on the horizontal axis of the preceding figure;

2. Algorithm, the name used by the respective ACE Challenge participant to refer to their algorithm;

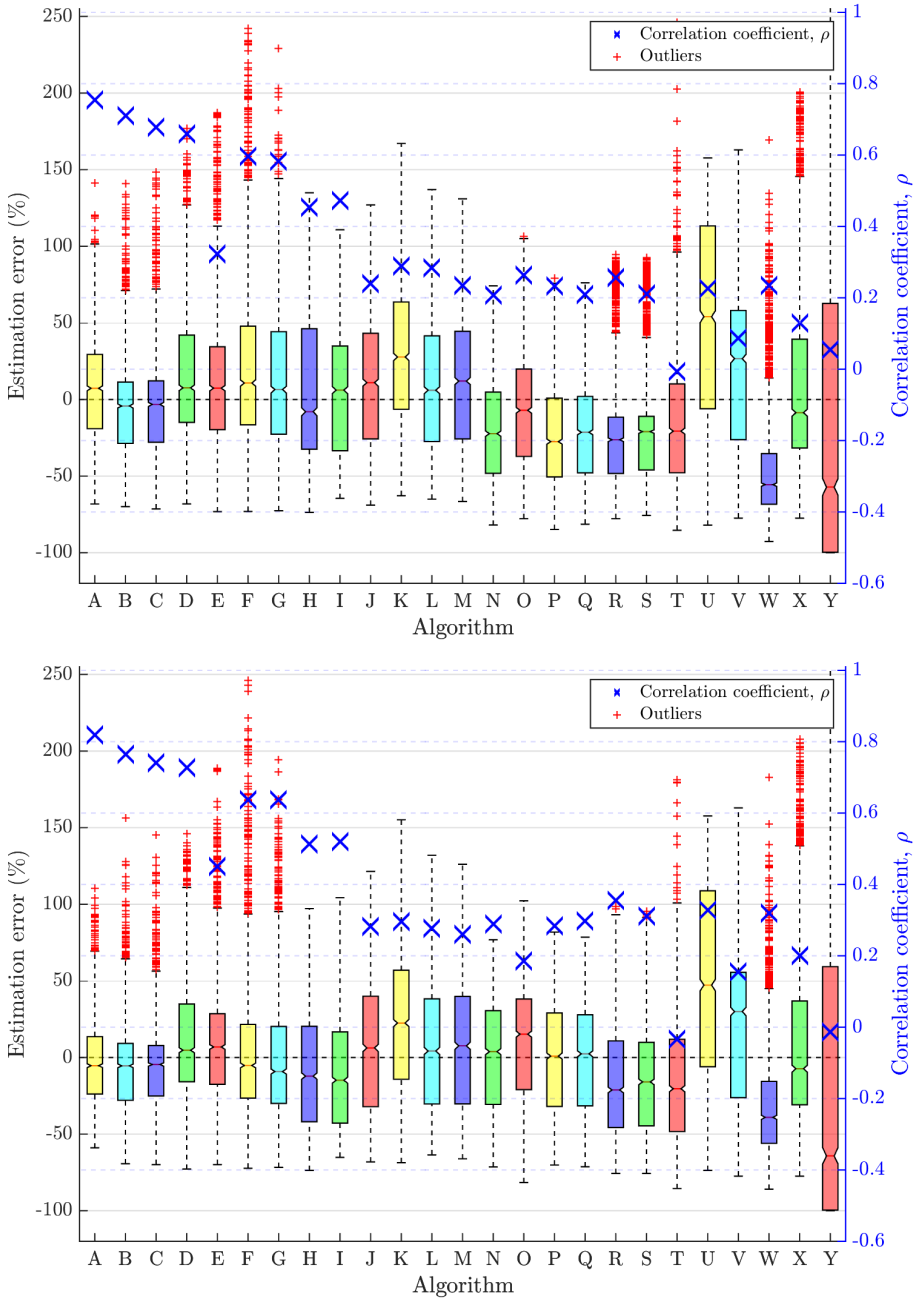

Figure 6.6: FB T_{60} estimation error in all noises and all SNRs for a) female talkers and b) male talkers

3. Mic. Config., the microphone configuration of the Evaluation dataset used to test the algorithm;

4. Bias, the mean error in the estimation results. Let $X = [x_0, x_1, \ldots x_{N-1}]$ equal the set of N ground truth T_{60} and DRR measurements, and let $\hat{X}$ equal the set of

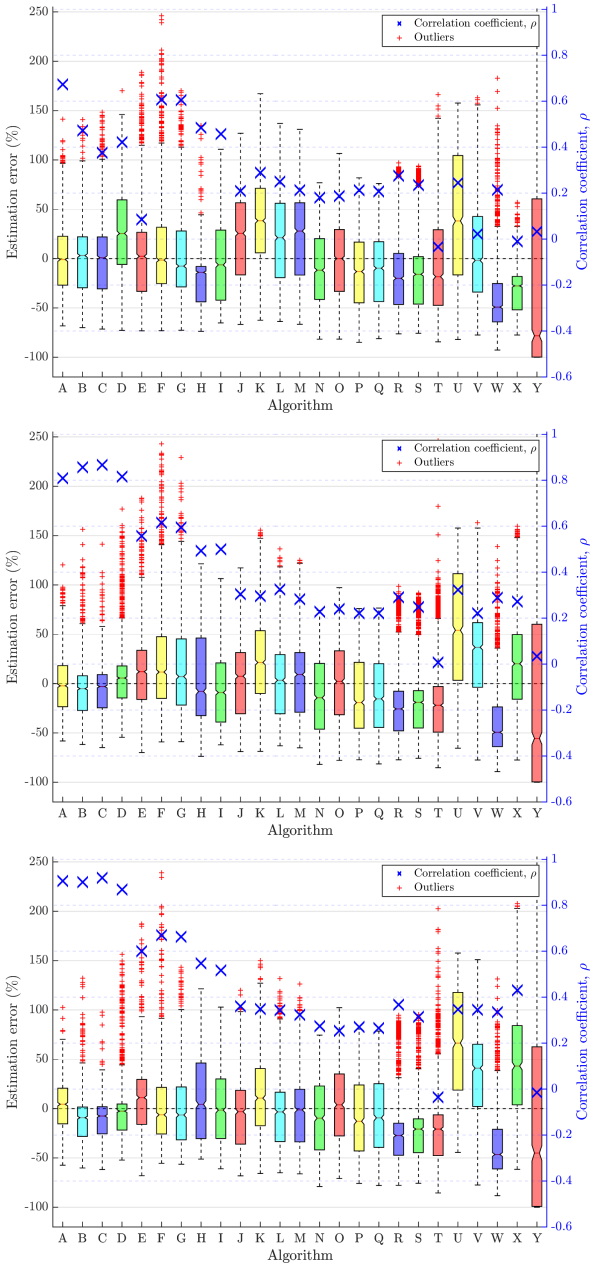

Figure 6.7: FB T_{60} estimation error in all noises and all SNRs for a) utterance length <5 s b) utterance length < 15 s and c) utterance length ⩾ 15 s

estimated results defined similarly. Then

$$\text{Bias} = \frac{1}{N}\sum_{n=0}^{N-1} \hat{x_n} - x_n. \tag{6.10}$$

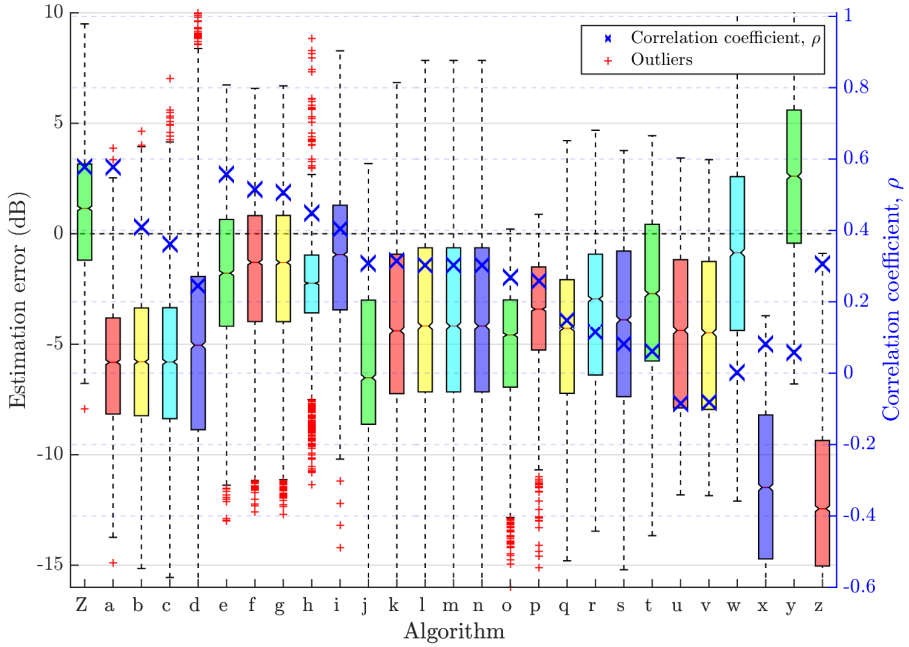

Figure 6.8: FB DRR estimation error in all noises for all SNRs

Ref.	Algorithm	Mic. Config.	Bias	MSE	ρ	RTF
Z	PSD est. in beamspace, bias comp. [182]	Mobile	1.07	8.14	0.577	0.757
a	PSD est. in beamspace (Raw) [182]	Mobile	-5.9	41.8	0.577	3.17
b	PSD est. in beamspace v2 [182]	Mobile	-5.7	43	0.41	0.844
c	PSD est. by twin BF [167]	Mobile	-5.71	44.9	0.362	0.614
d	Spatial Covariance in matrix mode [119]	Mobile	-5.37	61.2	0.244	0.627
e	NIRAv2 [176]	Single	-1.85	14.8	0.558	0.899
f	NIRAv3 [176]	Single	-1.62	14.7	0.515	0.899
g	NIRAv1 [176]	Single	-1.64	15	0.507	0.899
h	Particle velocity [183]	EM32	-2.38	10.4	0.449	0.134
i	Multi-layer perceptron [179]	Single	-1.14	15.9	0.405	0.0578
j	DENBE no noise reduction [5]	Chromebook	-6.04	51.2	0.308	0.0323
k	DENBE spectral subtraction [9]	Chromebook	-4.25	34.1	0.314	0.0589
l	DENBE spec. sub. Gerkmann [5]	Chromebook	-4.01	32.8	0.303	0.0477
m	DENBE filtered subbands [9]	Chromebook	-4.01	32.8	0.303	0.775
n	DENBE FFT derived subbands [9]	Chromebook	-4.01	32.8	0.303	0.0449
o	Normalized Overall SRMR (NOSRMR) Section 2.2. [180]	Chromebook	-5.1	34.3	0.269	1.04
p	Overall SRMR (OSRMR) Section 2.2. [180]	Chromebook	-3.71	20.6	0.259	0.829
q	NOSRMR Section 2.2. [180]	Mobile	-4.47	32	0.148	1.58
r	OSRMR Section 2.2. [180]	Mobile	-3.28	22.2	0.116	1.26
s	NOSRMR Section 2.2. [180]	Crucif	-4.05	31.1	0.0814	2.62
t	OSRMR Section 2.2. [180]	Crucif	-2.88	22.3	0.0616	2.09
u	NOSRMR Section 2.2. [180]	Single	-4.16	33.9	-0.0841	0.54
v	OSRMR Section 2.2. [180]	Single	-4.24	34.6	-0.0815	0.446
w	Per acoust. band SRMR Section 2.5. [180]	Single	-0.9	22.8	0.00192	0.578
x	Temporal dynamics [85]	Single	-11.4	147	0.0815	0.082
y	QA Reverb [177]	Single	2.51	23.6	0.0576	0.391
z	Blind est. of coherent-to-diffuse energy ratio [118]	Chromebook	-12.1	162	0.305	0.019

Table 6.5: DRR estimation algorithm performance in all noises for all SNRs

5. MSE, the mean squared error in the estimation results defined as

$$\text{MSE} = \frac{1}{N}\sum_{n=0}^{N-1}(\hat{x_n} - x_n)^2; \tag{6.11}$$

6. ρ, the Pearson correlation coefficient between the estimation results and the ground

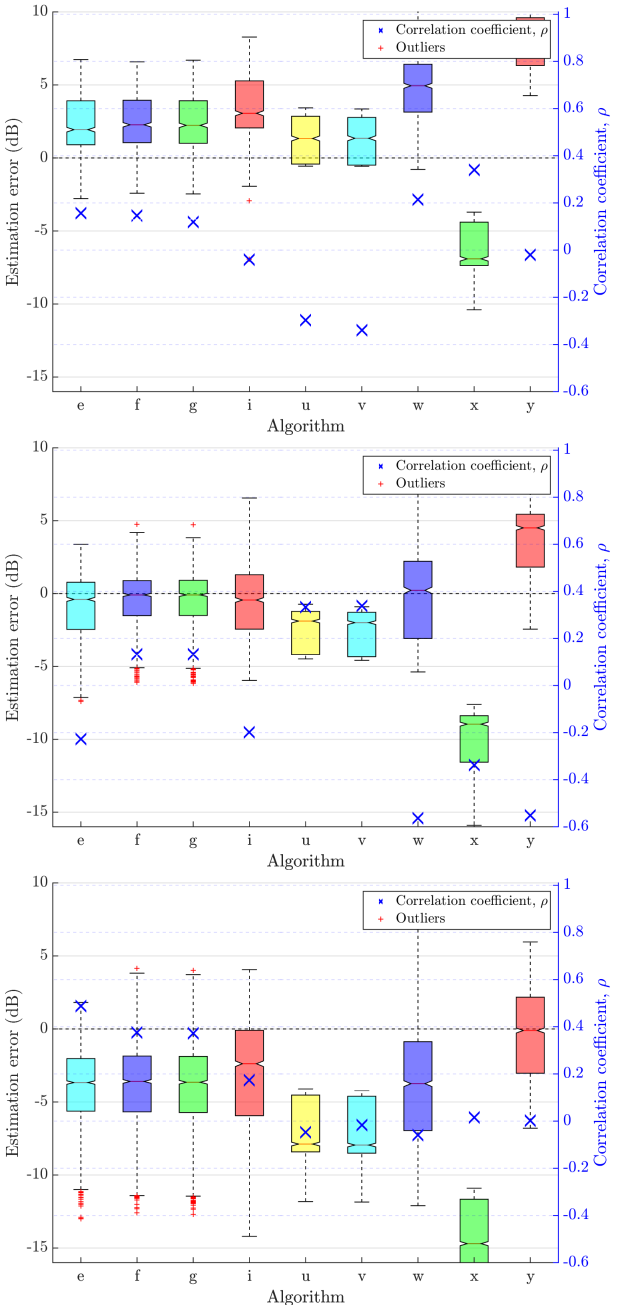

Figure 6.9: Single channel FB DRR estimation error in all noises and all SNRs for a) DRR <2 dB b) 2 ⩽DRR < 5 dB and c) DRR ⩾ 5 dB

truth defined as

$$\rho = \frac{E\{\hat{X}X\} - E\{\hat{X}\}E\{X\}}{\sqrt{(E\{\hat{X}^2\} - E\{\hat{X}\}^2)(E\{X^2\} - E\{X\}^2)}}; \tag{6.12}$$

7. RTF, the real-time factor, the mean computation time to perform the estimation task

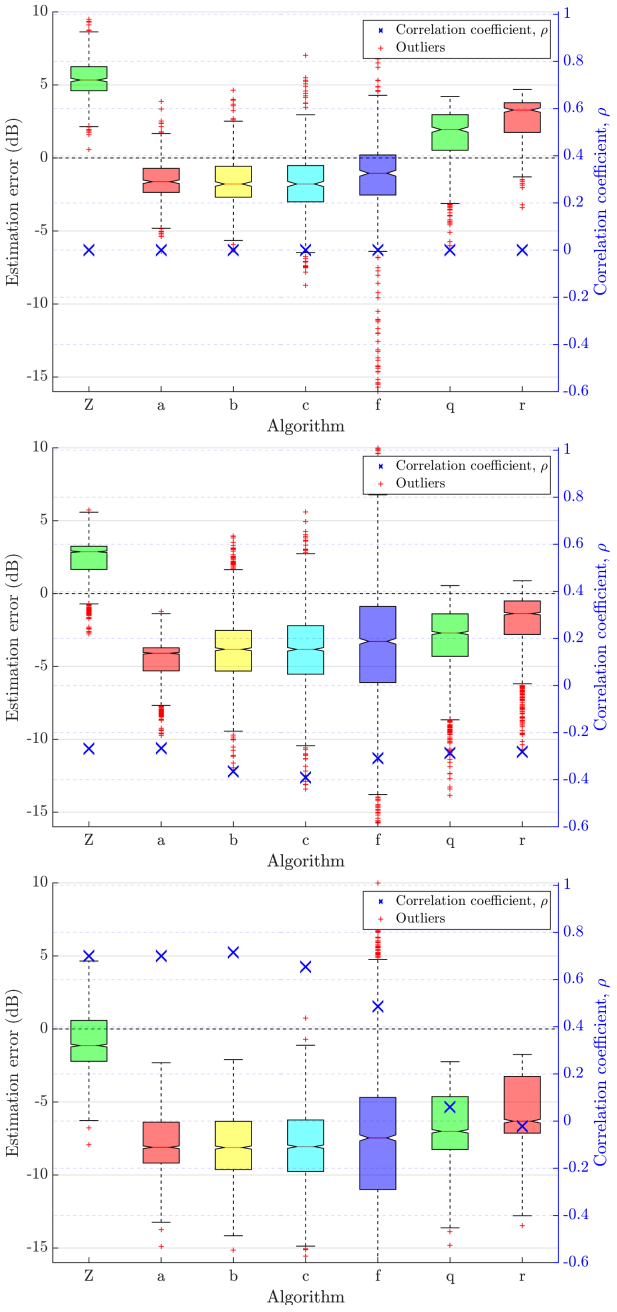

Figure 6.10: Mobile (3-channel) FB DRR estimation error in all noises and all SNRs for a) DRR <2 dB b) 2 ⩽DRR < 5 dB and c) DRR ⩾ 5 dB. Note that for b) there are strong negative correlations for all algorithms

over all relevant speech files.

By considering the bias, MSE, and ρ, it is possible to determine how well the estimator performs in these tests. For example, an estimator with a low bias and MSE might be giving an estimate close to the median for every speech file. By examining ρ, it will be

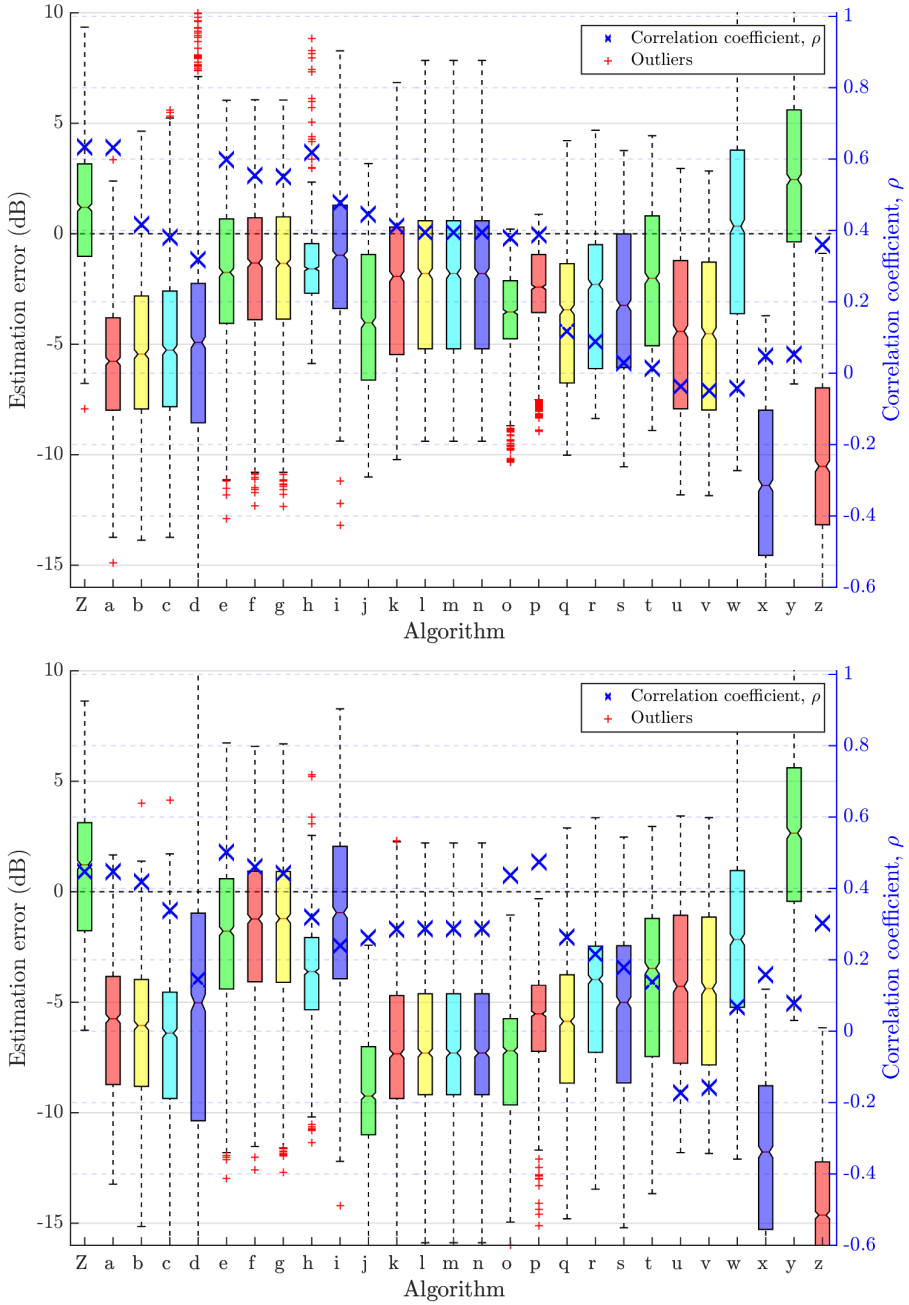

Figure 6.11: FB DRR estimation error in all noises at a), 18 dB **SNR and b)** −1 dB **SNR**

possible to distinguish between such an algorithm, which will have $\rho \to 0$, and a better algorithm which is more accurately estimating the parameter concerned which will have $\rho \to 1$. The RTF is useful for indicating whether the algorithm has potential to be used in practical applications that are constrained in computational complexity such as hearing aids and mobile devices.

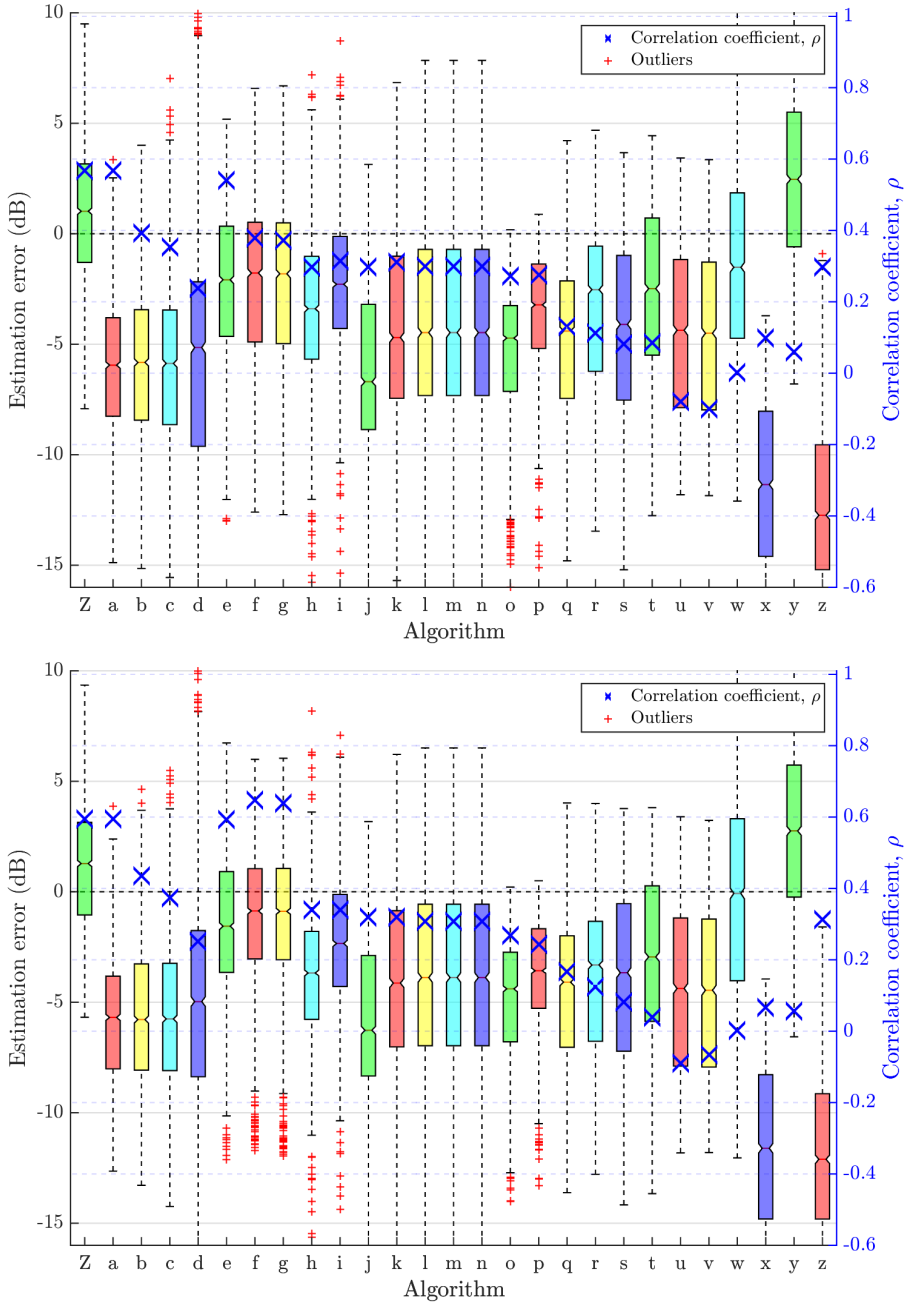

Figure 6.12: FB DRR estimation error in all noises and all SNRs for a) female talkers and b) male talkers

Fullband T_{60} estimation

The overall results for FB T_{60} estimation are shown in Fig. 6.3 and Table 6.4. There were no multi-channel approaches submitted that exploited the spatial information within the AIRs. The T_{60} is defined in the diffuse field where DRR < 0, and several algorithms rely on this

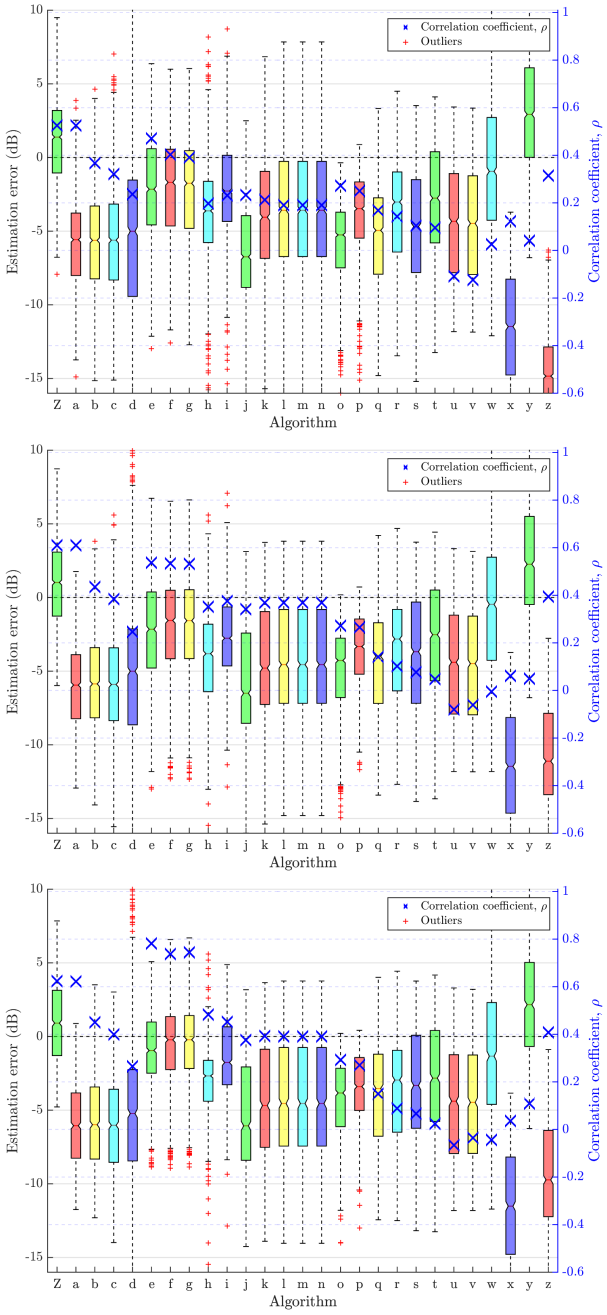

Figure 6.13: FB DRR estimation error in all noises and all SNRs for a) utterance length <5 s b) utterance length < 15 s and c) utterance length ⩾ 15 s

assumption. Further, the median DRR was approximately 5 dB, so the task was challenging. A recent approach using a null-steered beamformer to remove the direct path to deal with DRR ⩾ 0 was not submitted [90]. In spite of this the best performing algorithms achieved a correlation coefficient of $\rho \approx 0.8$. Around half the algorithms outperform the baselines from Gaubitch *et al.* [109], with the best performing showing a significant improvement which

suggests that non-intrusive T_{60} estimation is a mature field with estimators performing well. The most accurate algorithms are Single feature mapping or Analytical approaches. Features based on decay rates dominate the best approaches as expected since decay rates are a defining characteristic of a reverberant decay tail.

As can be seen in Fig. 6.4, most single channel T_{60} estimators underestimate longer T_{60}s. In some cases such as algorithms R, S and T this is a consequence of the Dev set containing a smaller range of T_{60}s than the Eval set, however in other cases such as F and G from Chapter 3 which were not tuned on the ACE corpus this is an inherent characteristic of the algorithm. Estimating T_{60} is more difficult for longer T_{60}s using decay rates as a feature because the gradients become very shallow and a small change in gradient at longer T_{60} with lead to a large change in the estimate.

As can be seen in Fig. 6.5, noise affects all algorithms. In general noise causes overestimation since the algorithms have difficulty distinguishing between noise and the noise-like tail of the reverberation. Machine learning approaches did not perform as well as the Single feature mapping or Analytical approaches. Modulation domain features worked well as features in mapping approaches but were not very effective as features in machine learning approaches. The blur kernel algorithms U and V, whilst showing promising results did not perform well in the ACE noise levels. The SDD algorithm reviewed in [109] is now the poorest performing algorithm, which is a satisfying observation in terms of scientific progress.

As can be seen in Fig. 6.6, most algorithms perform better in male speech. This effect is most pronounced for R, S, and T, which used a machine learning approach, as expected since the Dev set comprised only male talkers. The maximum likelihood and SDD-based algorithms are more accurate with longer utterances achieving $\rho \gtrsim 0.9$, as can be seen in Fig. 6.7.

Fullband DRR estimation results

The overall results for FB DRR estimation are shown in Fig. 6.8 and Table 6.5. Correlations are achieved up to $\rho \lesssim 0.6$, with the best performing algorithm exploiting spatial information in an analytical approach. The algorithms in general have large biases and MSEs. Also machine learning approaches are more common and successful than for the T_{60} task. This

suggests that the modelling and estimation of DRR is less well understood, and that there may be better features for DRR yet to be exploited by Analytical techniques, and that therefore this field of research is not as mature as T_{60} estimation.

Figures 6.9 and 6.10 show the results for DRR estimation for the Single and Mobile microphone configurations respectively for different ranges of DRR. These are shown separately since the ground truth DRR for the two microphones are different. Most algorithms overestimate at low DRR and underestimate at high DRR, suggesting that the algorithms are providing an estimate close to the middle of the range of DRRs, which is confirmed by the values of ρ, which show that there is poor correlation between the results and the ground truth.

As can be seen in Fig. 6.11, some algorithms are very susceptible to noise whilst for others it has little impact which reflects the diversity of techniques for DRR estimation. Figure 6.13 shows that utterance length has little impact on the DRR estimation algorithms unlike T_{60}. Similarly, as shown in Fig. 6.12, gender has little impact. This suggests that the statistics of speech are less useful for DRR estimation than for T_{60} estimation.

A frequency dependent T_{60} estimation algorithm and three frequency dependent DRR estimation algorithms were submitted. The results of these algorithms reflect the performance of the FB algorithms.

6.4.1 Test of the SDDSA-G and DENBE estimation algorithms on the ACE Corpus

The ACE Challenge provided an opportunity to evaluate the T_{60} and DRR estimation algorithms presented in Chapters 3 and 4. Whilst these are presented within the results above, further analysis is now performed.

Performance evaluation of the SDDSA-G T_{60} estimation algorithm

The performance evaluation was conducted using the Evaluation stage Single channel dataset and software provided by the ACE challenge.

Table 6.6 lists the algorithms to be compared along with their respective RTFs. For this test, SDD, SDD with Mel-spaced frequency bands (SDDMSB), and SDDSA-G, the

Label	Algorithm	Mic. Config.	RTF
W	SRMR Section 2.3. [180]	Single	0.457
X	Improved blind RTE [77]	Single	0.0269
Y	SDD [87]	Single	0.0224
α	SDDMSB [1]	Single	0.00435
G	SDDSA-G [1]	Single	0.0162
F	SDDSA-G retrained [8]	Single	0.0155

Table 6.6: T_{60} estimation algorithms for Figs. 6.14 to 6.16

proposed algorithm were implemented more efficiently using the Moore-Penrose [160] inverse and a matrix computation of the least-squares regression leading to significant reduction in computational complexity, but also significantly lower than the baseline algorithms, in particular SDDMSB. The algorithms W, X, Y, α, and G were discussed in Chapter 3. Also, algorithms W to C were the subject of the comparative evaluation in [109]. Algorithm F is identical to algorithm G excepting that the training procedure was performed using a larger range of T_{60} values. In algorithm G, T_{60} values from 0.2 to 0.95 s in 0.15 s steps, whereas in algorithm F, T_{60} values from 0.2 to 1.85 s in 0.15 s steps in anticipation of a potentially larger range of T_{60} values. The training was still based on simulated AIRs using the source-image method. Figures 6.14, 6.15 and 6.16 show the results for algorithms W, X, Y, α, G, and G in babble, ambient and fan noise at low (−1 dB), medium (12 dB) and high (18 dB) SNRs. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

The SDD algorithm exhibits a very high large bias at low SNRs. The SDD algorithm appears to exhibit unusual behaviour in ambient noise, and at low SNRs in fan noise. This is because the algorithm has a limit that prevents the T_{60} from being negative and thus for very high T_{60}s or large levels of noise the estimator returns a T_{60} of zero. The SDDMSB algorithm is more tolerant of noise than SDD, but that the SNR must be above 30 dB for performance to approach the proposed algorithm. Overall, the results show that the SDDSA-G algorithm and the retrained version outperform algorithms W, X, Y, α for most noise types and noise levels showing a smaller bias and smaller variance.

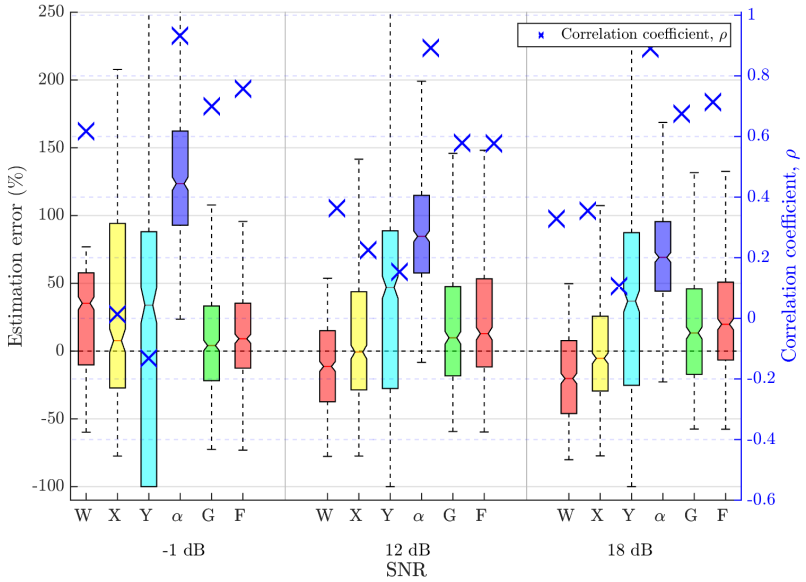

Figure 6.14: T_{60} **estimation error for babble noise for algorithms W, X, Y,** α**, G, and F for low** (-1 dB)**, medium** (12 dB) **and high** (18 dB) **SNRs**

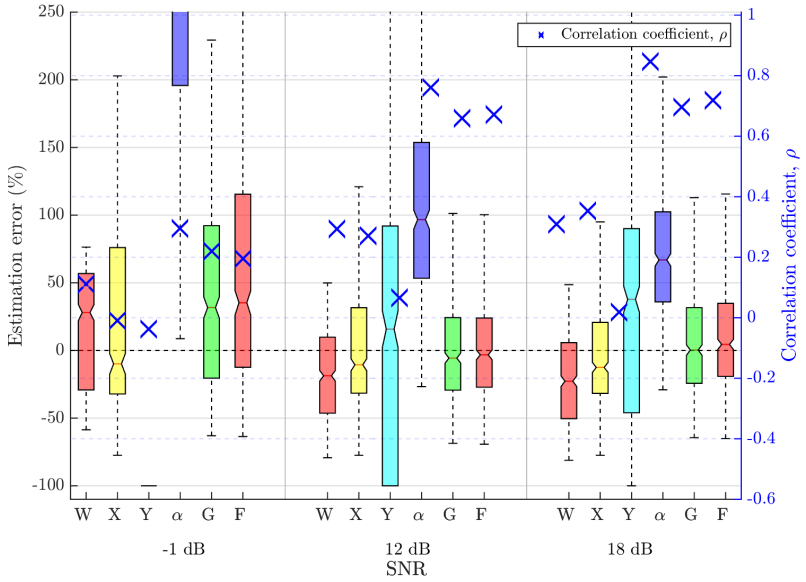

Figure 6.15: T_{60} **estimation error for fan noise for algorithms W, X, Y,** α**, G, and F for low** (-1 dB)**, medium** (12 dB) **and high** (18 dB) **SNRs**

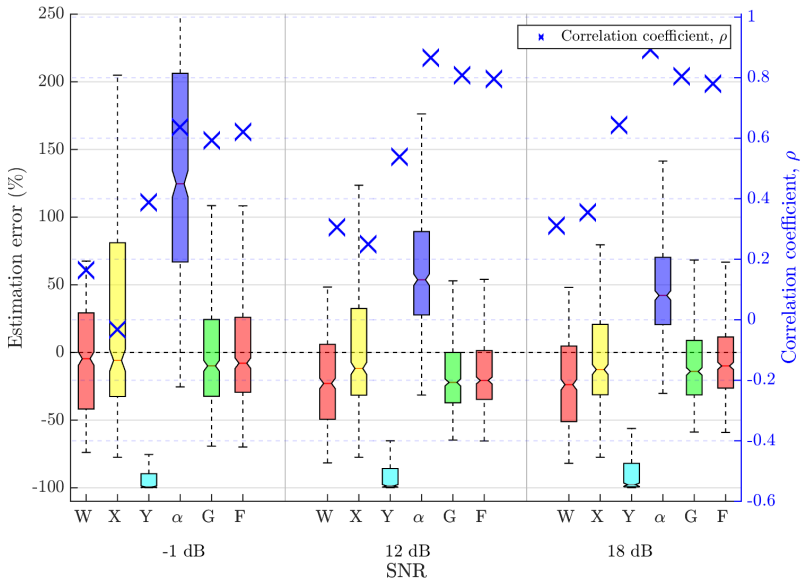

Figure 6.16: T_{60} **estimation error for ambient noise for algorithms W to F for low** (-1 dB)**, medium** (12 dB) **and high** (18 dB) **SNRs**

Performance evaluation of the DENBE DRR estimation algorithm

Table 6.7 lists the algorithms to be compared along with their respective RTFs. The DENBE algorithms (j to m) require the ground truth noise and the DoA. Since the ACE challenge is a fully blind test, ground truth noise and DoA were not available. Additional pre-processing was therefore performed resulting in several variants of the DENBE algorithm under test. In all variants time alignment of the two input channels was applied using *sigalign.m* [37]. In the first test denoted j, no noise reduction was applied. In the second test, denoted k, spectral subtraction was performed using minimum statistics [28, 37]. In the third test, denoted l, spectral subtraction was performed using Minimum Mean Squared Error (MMSE) power estimation [30, 37]. A further two tests of the algorithm in [5] were made. In the fifth test, denoted n, DRR estimates in ISO frequency bands were derived by transforming from the Short Time Fourier Transform (STFT) frequency bands in (4.22). In the sixth test, denoted m, DRR estimates in ISO frequency bands were obtained by filtering the input to the algorithm using an 8th order Butterworth filter and estimating the DRR using (4.23) over all frequencies. In addition, two baselines are used for comparison of the fullband DRR results, Falk *et al.* [85], and Jeub *et al.* [118] denoted x and z respectively.

Whilst the DRR estimator is computationally efficient, it is expected that variant m

Label	Algorithm	Mic. Config.	RTF
x	Temporal dynamics [85]	Single	0.0819
z	Blind est. of coherent-to-diffuse energy ratio [118]	Chromebook	0.019
j	DENBE no noise reduction [5]	Chromebook	0.0323
k	DENBE spectral subtraction [9]	Chromebook	0.0602
l	DENBE spec. sub. Gerkmann [5]	Chromebook	0.0474
n	DENBE FFT derived subbands [9]	Chromebook	0.0449
m	DENBE filtered subbands [9]	Chromebook	0.775

Table 6.7: DRR estimation algorithms for Figs. 6.17 to 6.19

will have a much higher RTF than the other versions since the algorithm must be called in every ISO frequency band up to half the sample rate. In addition to estimation performance, the RTF of each algorithm was compared. Algorithm x was tested using Matlab on an Intel Xeon X5675 processor with a clock speed of 3.07 GHz, whilst algorithms z, j, k, and l, n, and m were tested using Matlab on an Intel Xeon E5-2643 processor with a clock speed of 3.30 GHz. These two processors have similar levels of performance.

Results for algorithms x, z, j, k, and l in the three noise types, Ambient, Fan, and Babble are shown in Figs. 6.17, 6.18, and 6.19 respectively. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers. Outliers are not shown.

Algorithm j gives better performance than the baselines under all conditions. With the addition of noise reduction in algorithms k and l, the estimation performance is greatly improved for fan and ambient noise regardless of the method of noise reduction.

Results in ISO frequency bands for algorithms n and m in the three noise types at 18 dB SNRs are shown in Figs. 6.21 to 6.25. Algorithm n achieves some success at 200 Hz and above in ambient and fan noise, and above 1200 Hz in babble noise. Below 200 Hz the algorithm is is unable to estimate the DRR and is returning a value of −20 dB in most cases. As such the error is large and negative with a small variance. Algorithm m achieves better performance below around 250 Hz. This suggests that a combination of the two algorithms using the results of n above 200 Hz and the results of algorithm m below 200 Hz would be a better estimator than either estimator on its own. However, the variance in most frequency bands is large.

Table 6.7 shows the RTF for each algorithm computed by dividing the total CPU

time used to process all 4500 noisy reverberant speech files in the ACE Evaluation dataset by the combined length of all the noisy reverberant speech files. As expected, algorithm m has a much larger RTF than the other algorithms, however it would still be able to operate in real time in a practical application.

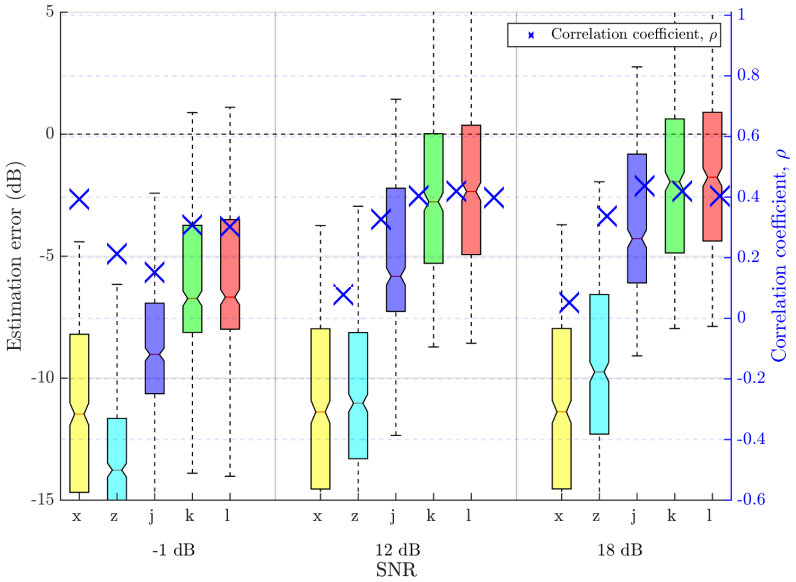

Figure 6.17: DRR estimation error for Ambient noise for algorithms x, z, j, k, and l for low (−1 dB), medium (12 dB) and high (18 dB) SNRs

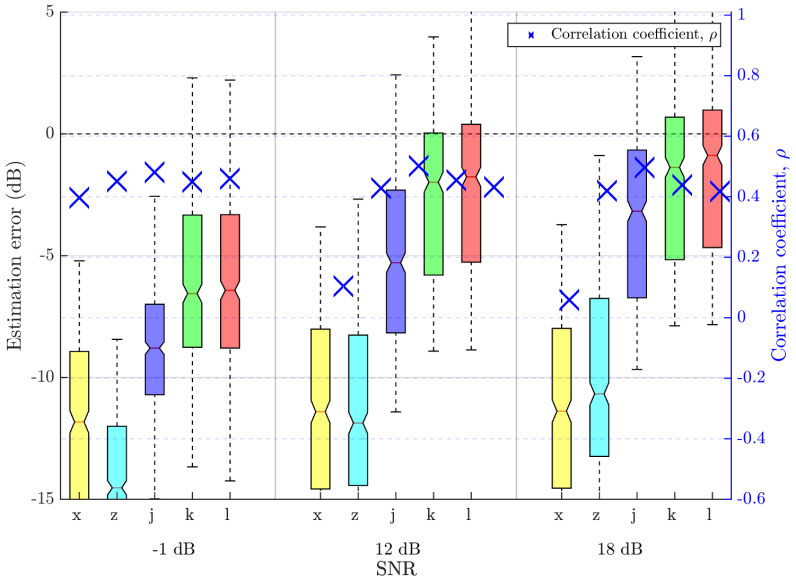

Figure 6.18: DRR estimation error for Fan noise for algorithms x, z, j, k, and l for low (−1 dB), medium (12 dB) and high (18 dB) SNRs

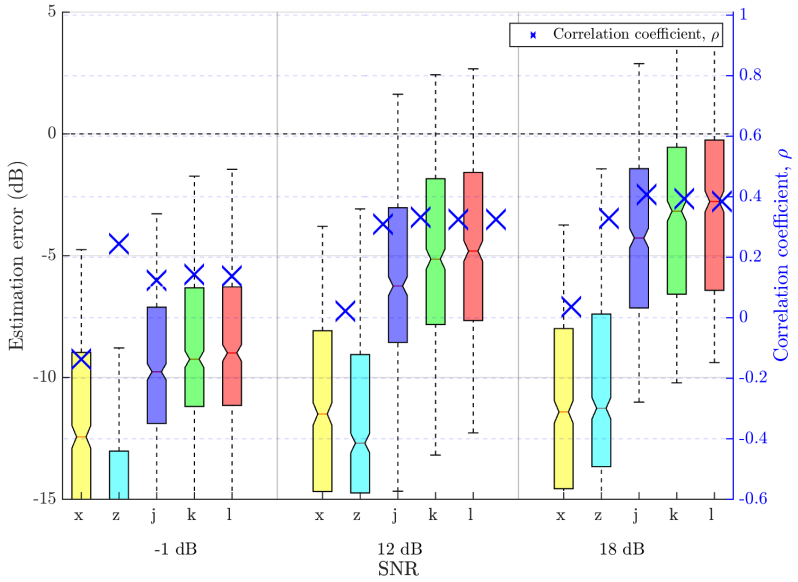

Figure 6.19: DRR estimation error for Babble noise for algorithms x, z, j, k, and l for low (−1 dB), medium (12 dB) and high (18 dB) SNRs

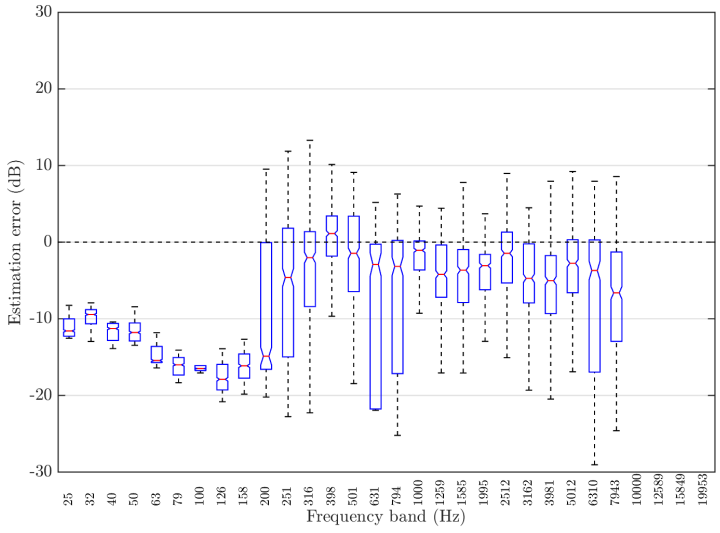

Figure 6.20: DRR estimation error for Ambient noise for algorithm n for high (18 dB) SNRs

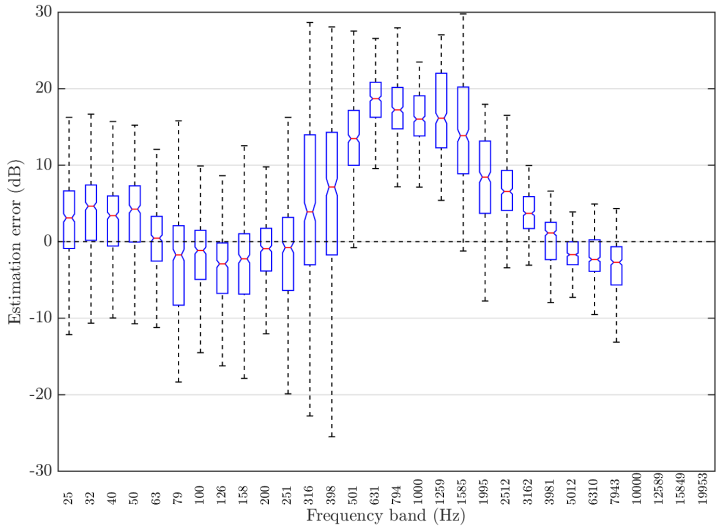

Figure 6.21: DRR estimation error for Ambient noise for algorithm m for high (18 dB) SNRs

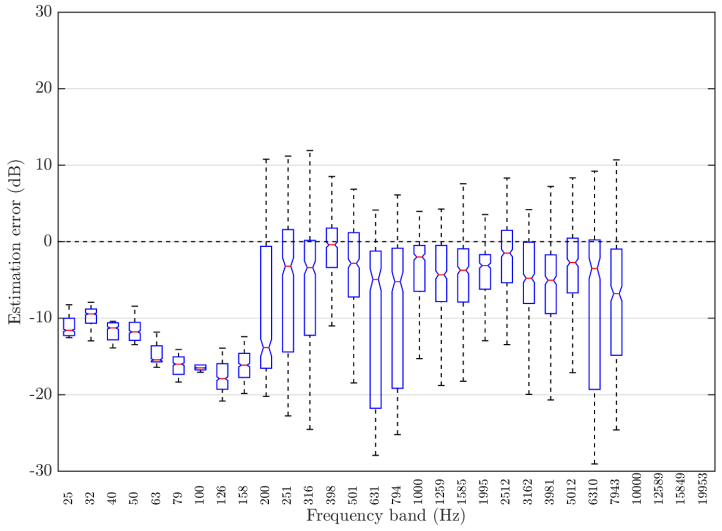

Figure 6.22: DRR estimation error for Fan noise for algorithm n for high (18 dB) SNRs

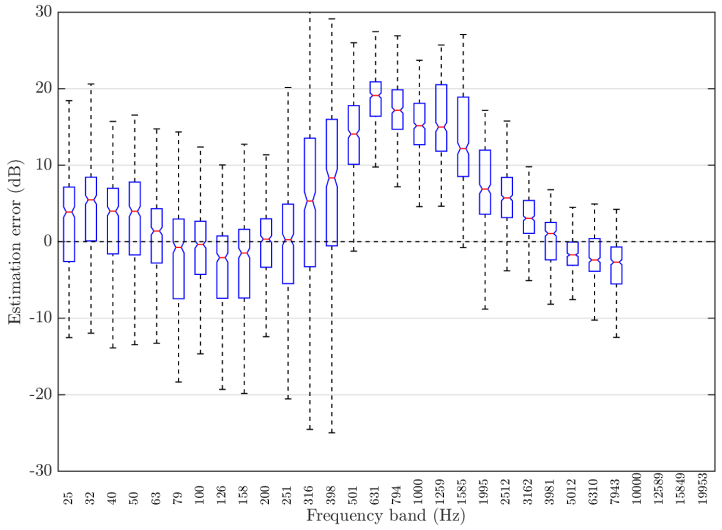

Figure 6.23: DRR estimation error for Fan noise for algorithm m for high (18 dB) SNRs

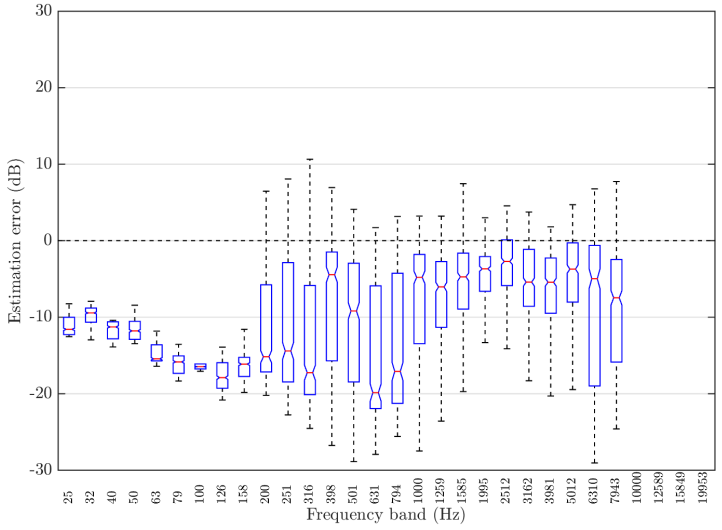

Figure 6.24: DRR estimation error for Babble noise for algorithm n for high (18 dB) SNRs

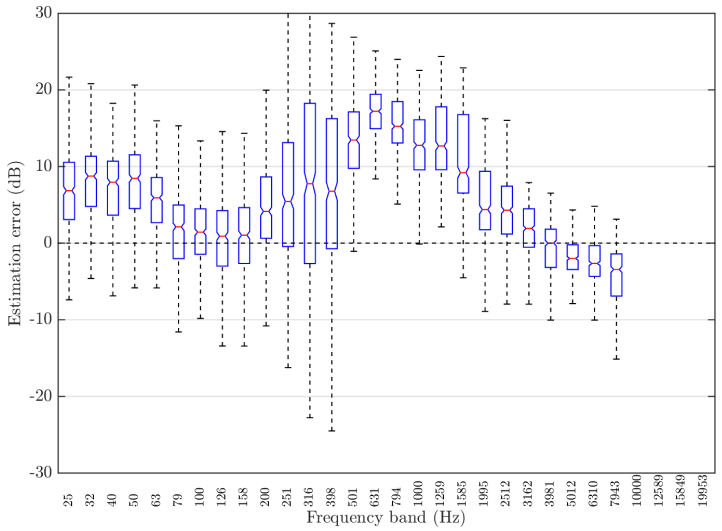

Figure 6.25: DRR estimation error for Babble noise for algorithm m for high (18 dB) SNRs

6.5 Conclusion

A corpus of multi-channel realistic noisy reverberant speech has been created, initially to support the ACE Challenge, but more widely to support any related research. Seven rooms are provided, each with two source-microphone configurations. For each of the 14 room-source-microphone configurations, 5 different microphone arrays are provided giving up to 50 channels of three different types of noise, along with their AIRs, and T_{60} and DRR measurements. The corpus allows researchers to construct noisy reverberant speech utterances with a range of T_{60}, DRRs, and noises at different SNRs using different microphone arrays in order to evaluate the performance of speech enhancement and speech recognition applications. The effect of occupancy on the T_{60} has also been demonstrated. The database will be freely available after the ACE Challenge has concluded.

The ACE Challenge has stimulated a significant research effort using a common data foundation which would not have happened without the creation of the corpus. In particular, DRR estimation algorithm "s" from Australian National University and T_{60} estimation algorithm "D" from Friedrich-Alexander University Erlangen-Nürnberg (FAU) were developed specifically for the ACE Challenge.

The results show that non-intrusive T_{60} estimation is a mature field, where the performance of blind estimators has significantly improved since the study in [109], and

further developments are likely to give marginal gains. Non-intrusive DRR estimation however is a significantly less mature field. Joint-estimation algorithms did not outperform algorithms focused on a single task, and their performance overall suggests leads to the conclusion that the main distinguishing features of T_{60} and DRR are different.

For non-intrusive T_{60} estimation, the current best-performing algorithm can estimate T_{60} to within an RMS error of $\approx 250\,\text{ms}$ and a $\rho \approx 0.8$ for typical operating scenarios of -2 to $18\,\text{dB}$ SNR. For non-intrusive DRR estimation, the best performing algorithm can estimate DRR to within an RMS error of $\approx 3\,\text{dB}$ and a $\rho \approx 0.6$ for typical operating scenarios of -2 to $18\,\text{dB}$ SNR.

The ACE Challenge has also provided an opportunity to evaluate the SDDSA-G and DENBE algorithms on real recordings of AIRs and noise in contrast to the previously reported performance which was on simulated data. The speech in the ACE corpus is free-running and less uniform than the TIMIT database used in [5], and so is representative of conversational speech rather than read speech. The estimation performance of the DENBE DRR estimator in fullband outperforms all existing algorithms at medium and high SNRs, and has a low RTF making it suitable for real-time applications. The performance in ISO frequency bands produced encouraging results based on small additions to the DENBE algorithm.

Chapter 7

Conclusion

THE starting point for this work was to determine whether it is possible to estimate acoustic parameters applicable to speech enhancement and speech recognition systems non-intrusively from single- and multi-channel speech degraded by additive, convolutive and non-linear degradations, in real-time or near real-time with sufficient accuracy to better enable speech enhancement.

In Chapter 2, existing methods for non-intrusively estimating the SNR for a speech utterance degraded by additive noise are reviewed. In the evaluation, estimates using existing methods are improved upon using a mapping function. The results show that the best performing algorithm with a mapping function can estimate the SNR in a 10 s speech utterance to within a spread of 7.3 dB for SNRs from −5 dB to 35 dB for 95% of the results. The algorithm has an RTF of 0.35 so the estimation can be performed in real-time.

In Chapter 3, a novel method for non-intrusively estimating T_{60} from speech degraded by the convolutive effects of reverberation, and additive noise is presented. The algorithm is based on the SDD method, which is highly biased in the presence of additive noise. By selecting the parts of the spectrum dominated by the speech, the method reduces the influence of the noise on the decay rates used to derive the T_{60} estimate, thus reducing the bias. The SDDSA-G algorithm provides increased accuracy and significant robustness to additive noise with Root Mean Square (RMS) T_{60} estimation errors of 120 ms or better for SNRs of 5 dB and above over a wide range of realistic additive noise degradations in real-time. The SDDSA-G algorithm requires an estimate of the SNR, and the best estimators

in Chapter 2 were shown to be sufficiently accurate to give results not significantly different from when the ground truth SNR is used.

In Chapter 4, a novel method for non-intrusively estimating the DRR from multi-channel speech degraded by additive noise and the convolutive effects of reverberation, the DENBE algorithm, is presented. This method determines the DRR using a null-steered beamformer. By steering a null at the source of the speech, the direct path is cancelled out leaving only the reverberation. Using the relative power of the reverberant speech signal, the power in the reverberation, and knowledge of the gain characteristics of the beamformer, the DRR is estimated. The DENBE algorithm was shown to provide estimates with a standard deviation of 4 dB over DRRs from −5 to 5 dB across a wide range of room sizes, reverberation times, and source-receiver distances.

The hitherto unexplored question of whether clipped samples can be detected in speech degraded by clipping, additive noise and the coding and decoding process of a communications channel has led to the development of a range of approaches in both time and frequency domain presented in Chapter 5. The algorithms outperform the optimized baseline, with F1 reaching 0.6 for Global System For Mobile Communications (GSM) 06.10 decoded speech with babble noise at 20 dB SNR, and clipping levels in the range −5 to −20 dBFS.

Alongside the algorithmic development, an international research challenge was run for which a novel noisy reverberant speech corpus was developed, which is presented in Chapter 6. The purpose of the ACE Challenge was to determine the state-of-the-art in T_{60} and DRR estimation, and stimulate new research. Nine research groups from eight different countries around the world participated with some developing new algorithms specifically as a result of the challenge. It can therefore be concluded therefore that the objectives were met. The ACE Challenge also provided the opportunity to test the SDDSA-G and DENBE algorithms presented in Chapters 3 and 4 on new data. The SDDSA-G T_{60} estimation algorithm presented in Chapter 3 demonstrated top-quartile performance both when used with simulated noise and AIRs, and when applied to the measured AIRs and matched noise recordings of the ACE Challenge. The implementation of the algorithm has been improved upon resulting in very low computational complexity and is available to download[1]. The

[1]Code available on http://www.commsp.ee.ic.ac.uk/∼sap/projects/blind-estimation-of-acoustic-parameters-from-speech/blind-t60-estimator/ password:bBs96K

DENBE DRR estimation algorithm presented in Chapter 4 demonstrated good performance in simulated data, and outperformed other 2-channel entries in the ACE Challenge.

This thesis set out to determine whether acoustic parameters could be estimated with sufficient accuracy to enable speech enhancement and recognition. It is encouraging therefore, that the SDDSA-G algorithm has been employed by several research groups as part of their speech enhancement and recognition algorithms [16, 184, 185], and that without additional training, the SDDSA-G algorithm performed well on the ACE Challenge Eval dataset, showing that it gives usable results under different conditions to those used for development of the algorithm. The DENBE algorithm performed well in the ACE Challenge and has a very low RTF making it well suited to practical implementations.

It can be concluded therefore that non-intrusive estimation of acoustic parameters from degraded speech is possible within the bounds set by this thesis.

7.1 What the future holds

At the start of this work, there was much focus on the T_{60} as a useful acoustic parameter, and work was ongoing to determine how accurately this could be estimated [109], which was also part of the inspiration for the ACE Challenge in Chapter 6. In May 2013, when the T_{60} estimation algorithm of Chapter 3 was presented at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) in Vancouver, Canada, this algorithm was state-of-the-art. The evidence presented in Chapter 6 shows that in the intervening years, T_{60} estimation has moved significantly further forward. As a result of the ACE Challenge, an accurate frequency dependent T_{60} estimation method has now been proposed [178]. The limitations of current methods means however that such great progress towards a perfect correlation coefficient approaching 1.0 is unlikely to be so rapid in the coming years, as this field is now relatively mature.

In the field of DRR estimation however, developments here are more recent and the value of this parameter is still gaining recognition. Prior to 2010, no methods had been presented for non-intrusive estimation of this parameter. Therefore improvements are expected to give correlations approaching 0.8 in the coming years, as has been seen in the field of T_{60} estimation as the time and frequency domain features that define this

parameter are better determined. Until then, machine learning approaches using many features are still likely to be amongst of the most successful methods, but this is ultimately an unsatisfactory approach from the point-of-view of truly understanding the acoustic properties of the reverberant sound field.

There are however, still some great difficulties in estimating even established acoustic parameters. For example T_{60} must be measured in the diffuse field beyond the critical distance. When estimating non-intrusively, it may be impossible to determine whether the observation is beyond the critical distance or not. For DRR, the energy in the direct path must be determined, which is not possible with any accuracy in practice, and there is no single established method for estimating this. It has also not been established with any certainty whether DRR estimation is actually theoretically possible from a single channel speech signal for DRRs less than 0 dB. Also there is an interdependence between T_{60} and DRR. With a high DRR even a very large T_{60} is unlikely to adversely affect intelligibility, whilst with a low DRR, even low values of T_{60} can dramatically colour the observation.

Very recently, focus has been shifting towards the Clarity Index (C_{50}), which is defined as the ratio of the energy in the early reflections over the energy in the late reflections [111]. This parameter compares the energy in the AIR up until 50 ms from the arrival of the direct signal with the energy in the remainder of the signal, and therefore circumvents some of the problems associated with established acoustic parameters. Since early reflections prior to 50 ms tend to only spectrally colour a signal, whilst late reflections cause temporal smearing, a high C_{50} will correspond to a signal with higher quality and intelligibility [122]. This parameter has also been shown to be highly correlated with Automatic Speech Recognition (ASR) performance [186]. Whilst a second phase of the ACE Challenge was held to allow new participants to submit algorithms specifically developed for the challenge, C_{50} unfortunately fell outside of the scope of the work. Were the ACE Challenge to be held again today, C_{50} would undoubtedly be included as one of the estimation tasks. This raises the question of what new acoustic parameters may emerge in the future and what improvements in speech enhancement might be made as a result of those parameters.

Much research work still relies on the LTI assumption, and in this thesis this was an underlying assumption in Chapters 3, 4, and 6. In practice, more complex degradations

may be encountered which challenge this assumption. A person talking in a room will move around the room even if they are trying to keep still, and heat sources and draughts in the room will cause air currents which will change the acoustics continuously. The talker may even leave the room altogether during a conversation and enter another acoustic environment, and this may occur several times during a conversation. Also, communications devices are not perfectly linear, and recovery from overload conditions in such devices will take a finite time period. Estimating acoustic parameters under such challenging - and realistic - circumstances is the subject for further work.

Bibliography

[1] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

[2] J. Eaton, M. Brookes, and P. A. Naylor, "A comparison of Non-Intrusive SNR estimation algorithms and the use of mapping functions," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[3] J. Eaton and P. A. Naylor, "Detection of clipping in perceptually encoded speech signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[4] ——, "Noise-robust detection of peak-clipping in decoded speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 7019–7023.

[5] J. Eaton, A. H. Moore, P. A. Naylor, and J. Skoglund, "Direct-to-reverberant ratio estimation using a null-steered beamformer," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 46–50.

[6] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE Challenge - corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

[7] ——, "Proceedings of the ACE Challenge Workshop, a satellite event of IEEE-WASPAA," New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1510.00383

[8] J. Eaton and P. A. Naylor, "Reverberation time estimation on the ACE corpus using the SDD method," in *Proc. ACE Challenge Workshop, a satellite event of IEEE-WASPAA*, New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1510.01193

[9] ——, "Direct-to-reverberant ratio estimation on the ACE corpus using a two-channel beamformer," in *Proc. ACE Challenge Workshop, a satellite event of IEEE-WASPAA*, New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1510.07546

[10] T. Neeld, J. Eaton, P. A. Naylor, and D. Shipworth, "A novel method of determining events in combination gas boilers: Assessing the feasibility of a single-point acoustic sensor," *Building and Environment*, vol. 100, pp. 1–9, May 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360132316300361

[11] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustics parameters: The ACE challenge," *IEEE Trans. Audio, Speech, Lang. Process.*, Dec. Submitted.

[12] ——, "ACE Challenge results technical report," Imperial College London, Technical Report, 2016. [Online]. Available: http://www.commsp.ee.ic.ac.uk/~sap/uploads/data/ACE/ACE_technical_report.pdf

[13] Ofcom, "Communications market report," Aug. 2014. [Online]. Available: http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr14/2014_UK_CMR.pdf

[14] ——, "International communications market report," Dec. 2014. [Online]. Available: http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr14/icmr/ICMR_2014.pdf

[15] C. J. Wild, A. Yusuf, D. E. Wilson, J. E. Peelle, M. H. Davis, and I. S. Johnsrude, "Effortful listening: the processing of degraded speech depends critically on attention," *J. of Neuroscience*, vol. 32, no. 40, pp. 14 010–14 021, Oct. 2012.

[16] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *Proc.*

REVERB Challenge Workshop, vol. 1, 2014, pp. 1–8. [Online]. Available: http://reverb2014.dereverberation.com/workshop/reverb2014-papers/1569898243.pdf

[17] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007. [Online]. Available: http://alexandria.tue.nl/extra2/200710970.pdf

[18] M. J. Harvilla and R. M. Stern, "Least squares signal declipping for robust speech recognition," in *Proc. Interspeech Conf.*, vol. 1, Singapore, Sep. 2014, pp. 2073–2077.

[19] 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.

[20] 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.

[21] Soundjay.com, "Sound effects website," http://www.soundjay.com/.

[22] P. A. Naylor and N. D. Gaubitch, "Acoustic signal processing in noise: It's not getting any quieter," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, 2012, pp. 1–6.

[23] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb 2012.

[24] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, Mar. 2005, pp. 953–956.

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

[26] M. Grimm and K. Kroschel, Eds., *Robust Speech Recognition and Understanding*. I-Tech, Vienna, Austria, 2007.

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

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

[29] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 4266–4269.

[30] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[31] NIST, "NIST speech tools," http://www.nist.gov/speech/tools, 2006.

[32] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech Conf.*, Sep. 2008, pp. 2598–2601.

[33] A. Narayanan and D. L. Wang, "A CASA-based system for long-term SNR estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2518–2527, Nov. 2012.

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

[35] P. Kabal, "Measuring speech activity," McGill Univ, Tech. Rep., Aug. 1999.

[36] R. W. Berry, "Speech-volume measurements on telephone circuits," *Proc IEE*, vol. 118, no. 2, pp. 335–338, Feb. 1971.

[37] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997–2015.

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

[39] D. Mauler and R. Martin, "Noise power spectral density estimation on highly correlated data," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC)*, Sep. 2006, pp. 1–4.

[40] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1112–1123, 2008.

[41] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Workshop on Automatic Speech Recognition*, Paris, Sep. 2000, pp. 181–188.

[42] J. Meyer, K. U. Simmer, and K. D. Kammeyer, "Comparison of one- and two-channel noise-estimation techniques," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC)*, vol. 1, London, United Kingdom, Sep. 1997, pp. 137–145.

[43] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Process. Lett.*, vol. 11, no. 9, pp. 725–728, Sep. 2004.

[44] M. Vondrasek and P. Pollak, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," *Radio Engineering*, vol. 14, pp. 6–11, Jan. 2005.

[45] P. Pollák, "Speech signals database creation for speech recognition and speech enchancement applications," Associate Professor Inaugural Dissertation, Czech Technical University in Prague, Fac. of Elec. Eng., Prague, Czech Republic, 2002.

[46] D. Iskra, B. Groβkopf, K. Marasek, H. v. d. Heuvel, F. Diehl, and A. Kiessling, "SpeeCon - speech data for consumer devices: Database specification and validation," in *Proc. Intl. Conf. on Lang. Resources and Evaluation (LREC)*, no. 3, 2002, pp. 329–333.

[47] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 2006, pp. 153–156.

[48] F. Beritelli, S. Casale, R. Grasso, and A. Spadaccini, "Performance evaluation of SNR estimation methods in forensic speaker recognition," in *Intl. Conf. on Emerging Security Inf. Syst. and Tech. (SECURWARE)*, no. 4, Jul. 2010, pp. 88–92.

[49] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1988, pp. 651–654.

[50] J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4640–4643.

[51] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.

[52] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany, Apr. 1997, pp. 21–24.

[53] B. Nimens, "Sound ideas: sound effects collection," http://www.sound-ideas.com/6000.html.

[54] ITU-T, *Test signals for use in telephonometry*, Intl. Telecommunications Union (ITU-T) Recommendation P.501, Aug. 1996.

[55] ——, *Transmission Standards Recommendation P.48 Specification for an intermediate reference system*, International Telecommunications Union (ITU-T) Recommendation P.48, Nov. 1988.

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

[57] W. C. Sabine, *Collected Papers on acoustics (Originally 1921)*. Peninsula Publishing, 1993.

[58] C. F. Eyring, "Reverberation time in 'dead' rooms," *J. Acoust. Soc. Am.*, vol. 1, no. 2A, p. 168, 1930.

[59] ISO, *ISO-266 Acoustics - Preferred frequencies*, Intl. Org. for Standardization (ISO) Recommendation ISO-266, Mar. 1997.

[60] L. Couvreur and C. Couvreur, "On the use of artificial reverberation for ASR in highly reverberant environments," in *Proc. IEEE Benelux Signal Processing Symposium (SPS)*, no. 2, Hilvarenbeek, The Netherlands, Mar. 2000, pp. S001–S004.

[61] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[62] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

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

[64] ISO, *ISO-3382 Acoustics - Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters*, Intl. Org. for Standardization (ISO) Recommendation ISO-3382, May 2009.

[65] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. Audio Eng. Soc. (AES) Convention*, no. 108, Feb 2000, pp. 1–23. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10211

[66] ——, "Advancements in impulse response measurements by sine sweeps," in *Proc. Audio Eng. Soc. (AES) Convention*, no. 123, Vienna, May 2007, pp. 1–21.

[67] S. Mosayyebpour, H. Sheikhzadeh, T. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based inverse filtering of the room impulse response," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1617–1632, Jul. 2012.

[68] A. Baskind and O. Warusfel, "Methods for blind computational estimation of perceptual attributes of room acoustics," in *Proc. Audio Eng. Soc. (AES) Conf. on Virtual, Synthetic & Entertainment Audio*, no. 22, Jun. 2002, pp. 402–411.

[69] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, Jr., C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[70] R. Ratnam, D. L. Jones, and W. D. O'Brien, Jr., "Fast algorithms for blind estimation of reverberation time," *IEEE Signal Process. Lett.*, vol. 11, no. 6, pp. 537–540, Jun. 2004.

[71] J. Vieira, "Automatic estimation of reverberation time," in *Proc. Audio Eng. Soc. (AES) Convention*, no. 116, Berlin, Germany, May 2004. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12785

[72] ——, "Estimation of reverberation time without test signals," in *Proc. Audio Eng. Soc. (AES) Convention*, no. 118, Barcelona, Spain, May 2005. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=13215

[73] S. Vesa and A. Harma, "Automatic estimation of reverberation time from binaural signals," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Mar. 2005, pp. iii/281–iii/284.

[74] Y. Zhang, J. A. Chambers, F. F. Li, P. Kendrick, and T. J. Cox, "Blind estimation of reverberation time in occupied rooms," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Sep. 2006, pp. 1–5.

[75] H. W. Löllmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC)*, Sep. 2008, pp. 1–4.

[76] ——, "Reverberation time estimation for speech processing applications," in *Proc. Intl. Conf. on Acoustics (DAGA)*, Rotterdam, Netherlands, Mar. 2009, pp. 1655–1658.

[77] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC)*, Tel-Aviv, Israel, Aug. 2010, pp. 1–4.

[78] T. de M. Prego, A. A. de Lima, S. L. Netto, B. Lee, A. Said, R. W. Schafer, and T. Kalker, "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2811–2816, Apr. 2012.

[79] P. Kendrick, N. Shiers, R. Conetta, T. J. Cox, B. M. Shield, and C. Mydlarz, "Blind estimation of reverberation time in classrooms and hospital wards," *Applied Acoustics*, vol. 73, no. 8, pp. 770–780, Aug. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003682X12000357

[80] T. Jan and W. Wang, "Blind reverberation time estimation based on laplace distribution," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 2050–2054.

[81] S. Diether, L. Bruderer, A. Streich, and H.-A. Loeliger, "Efficient blind estimation of subband reverberation time from speech in non-diffuse environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Brisbane, Australia, May 2015, pp. 743–747.

[82] M. Wu and D. Wang, "A Pitch-Based Method for the Estimation of Short Reverberation Time," *Acta Acustica united with Acustica*, vol. 92, no. 2, pp. 337–339, Apr. 2006.

[83] M. Unoki and S. Hiramatsu, "Blind estimation method of reverberation time based on concept of modulation transfer function," in *Proc. Intl. Conf. on Acoustics (DAGA)*, Paris, France, Jun. 2008, pp. 4492–4496.

[84] A. Kurematsu, K. Takeda, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," in *European Speeech Commun. Assoc. (ESCA) Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases (SIOA)*, vol. 2, Noordwijkerhout, The Netherlands, Sep. 1989, pp. 43–46.

[85] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, 2010.

[86] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[87] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, Apr. 2008, pp. 329–332.

[88] N. Lopez, Y. Grenier, and I. Bourmeyster, "Low variance blind estimation of the reverberation time," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[89] R. Talmon and E. A. P. Habets, "Blind reverberation time estimation by intrinsic modeling of reverberant speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 156–160.

[90] B. Dumortier and E. Vincent, "Blind RT60 estimation robust across room sizes and source distances," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 5187–5191.

[91] F. Lim, M. R. P. Thomas, and I. J. Tashev, "Blur kernel estimation approach to blind reverberation time estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 41–45.

[92] F. Lim, M. R. P. Thomas, P. A. Naylor, and I. J. Tashev, "Acoustic blur kernel with sliding window for blind estimation of reverberation time," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

[93] T. J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *J. Audio Eng. Soc. (AES)*, vol. 49, no. 4, pp. 219–230, 2001.

[94] R. M. Cox, G. C. Alexander, and C. Gilmore, "Development of the connected speech test (CST)," *Ear and Hearing, J. of American Auditory Soc. (AAS)*, vol. 8, no. 5, pp. 119S–126S, 1987.

[95] J. D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, France, 1988.

[96] Y. Zhang, J. A. Chambers, P. Kendrick, T. J. Cox, and F. F. Li, "Acoustic parameter extraction from occupied rooms utilizing blind source separation," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, B. Gabrys, R. J. Howlett, and L. C. Jain, Eds.

Springer Berlin Heidelberg, 2006, vol. 4253, pp. 1208–1215. [Online]. Available: http://dx.doi.org/10.1007/11893011_153

[97] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[98] H. G. Hirsch, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. Interspeech Conf.*, Lisboa, Portugal, Sep. 2005, pp. 2697–2700.

[99] ITU-T, *Software Tools for Speech and Audio Coding Standardization*, International Telecommunications Union (ITU-T) Recommendation G.191, Sep. 2005.

[100] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Assoc. for Computational Linguistics (ACL) Intl. Workshop on Speech and Natural Language*, ser. HLT '91, Stroudsburg, PA, USA, 1992, pp. 357–362.

[101] J. Y. C. Wen and P. A. Naylor, "Semantic coloration space investigation: Controlled coloration in the bark-sone domains," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2007, pp. 311–314.

[102] J. Y. C. 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 and Noise Control (IWAENC)*, Paris, France, Sep. 2006, pp. 1–4.

[103] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Sydney, Australia, Jul. 2009, pp. 1–4.

[104] A. A. de Lima, F. P. Freeland, P. A. A. Esquef, L. W. P. Biscainho, B. C. Bispo, R. A. de Jesus, S. L. Netto, R. Schafer, A. Said, B. Lee, and A. Kalker, "Reverberation assessment in audioband speech signals for telepresence systems," in *Proc. Intl. Conf. on Signal Processing in Multimedia Applications (SIGMAP)*, vol. 1, Porto, Portugal, 2008, pp. 257–262.

[105] J. Kominek and A. W. Black, "The CMU Arctic databases for speech synthesis," in *Proc. Intl. Speech Commun. Assoc. (ISCA) Workshop on Speech Synthesis*, Pittsburgh, PA, USA, Jun. 2004, pp. 223–224.

[106] E. Lehmann and A. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1429–1439, Aug. 2010.

[107] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006. [Online]. Available: http://link.aip.org/link/?JAS/120/2421/1

[108] S. Shelley and D. Murphy, "Openair: An interactive auralization web resource and database," in *Proc. Audio Eng. Soc. (AES) Convention*, San Francisco, CA, USA, Nov. 2010. [Online]. Available: http://www.openairlib.net/

[109] N. D. Gaubitch, H. W. Löllmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012, pp. 1–4.

[110] E. A. P. Habets. (2003) Room impulse response generator for MATLAB. http://home.tiscali.nl/ehabets/rir_generator.html. [Online]. Available: http://home.tiscali.nl/ehabets/rir_generator.html

[111] H. Kuttruff, *Room Acoustics*, 4th ed. London: Taylor & Francis, 2000.

[112] L. R. Rabiner and R. W. Schafer, Eds., *Theory and Applications of Digital Signal Processing*. Pearson, 2010.

[113] I.-J. Bienaymé, "Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés," *Comptes rendus des Séances de l'Académie des Sciences*, vol. 37, pp. 309–324, Aug. 1853.

[114] I. R. Titze, *Principles of Voice Production*. Prentice Hall, 1994.

[115] R. J. Baken, *Clinical Measurement of Speech and Voice*. London, UK: Taylor & Francis Ltd., 1987.

[116] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, May 2005.

[117] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1793–1805, Sep. 2010.

[118] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Spain, 2011, pp. 1347–1351.

[119] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using d/r spatial correlation matrix model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2374–2384, Nov 2011.

[120] M. Kuster, "Estimating the direct-to-reverberant energy ratio from the coherence between coincident pressure and particle velocity," *J. Acoust. Soc. Am.*, vol. 130, no. 6, pp. 3781–3787, 2011. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/130/6/10.1121/1.3658446

[121] O. Thiergart, G. Del Galdo, and E. A. P. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *J. Acoust. Soc. Am.*, vol. 132, no. 4, pp. 2337–2346, 2012.

[122] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, Apr. 2010.

[123] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Springer, 2008.

[124] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[125] J. Gruber and L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems," *IEEE Trans. Commun.*, vol. 33, no. 8, pp. 305–307, Jan. 1985.

[126] A. Koski, "Rich recording technology," Nokia Product Engineering, Technical overall description, 2012. [Online]. Available: http://i.nokia.com/blob/view/-/1696152/data/2/-/Download2.pdf

[127] H. R. Pfitzinger and C. Kaernbach, "Amplitude and amplitude variation of emotional speech," in *Proc. Interspeech Conf.*, vol. 1, Brisbane, Australia, Sep. 2008, pp. 1036–1039.

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

[129] 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.

[130] ETSI, *GSM 06.10: European digital cellular telecommunications system (Phase 2); Full rate speech transcoding*, European Telecommunications Standards Institute (ETSI) ETSI Standard ES 300 580-2, Sep. 1994.

[131] ——, *Adaptive Multi-Rate (AMR) Speech Transcoding*, European Telecommunications Standards Institute (ETSI) ETSI Standard GSM 06.90, 1998.

[132] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, 2006.

[133] ITU-T, *Test signals and metering to be used on international sound programme connections*, International Telecommunications Union (ITU-T) Recommendation BS. 645–2, Mar. 1992.

[134] EBU, *Alignment level in digital audio production equipment and in digital audio recorders*, European Broadcasting Union (EBU) Recommendation R68-2000, 2000.

[135] SMPTE, *Reference Level for Digital Audio Systems*, Society of Motion Picture and Television Engineers(SMPTE) Recommendation R155-2004, Aug. 2004.

[136] G. Spikofski and S. Klar, "Levelling and loudness in radio and television broadcasting," Institut für Rundfunktechnik GmbH (IRT), EBU Technical Review, Jan. 2004.

[137] M. Paez and T. Glisson, "Minimum mean-squared-error quantization in speech PCM and DPCM systems," *IEEE Trans. Commun.*, vol. 20, no. 2, pp. 225–230, Apr. 1972.

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

[139] R. A. McDonald, "Signal-to-noise and idle channel performance of differential pulse code modulation systems - particular applications to voice signals," *Bell Syst. Tech. J.*, vol. 45, no. 1, pp. 1123–1151, 1966.

[140] F. Vilbig, "An analysis of clipped speech," *J. Acoust. Soc. Am.*, vol. 27, no. 1, pp. 207–207, 1955.

[141] H. L. F. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, 2nd ed. Dover Publications, Inc., 1885 (1954).

[142] A. Gallardo-Antolin, F. D. de Maria, and F. Valverde-Albacete, "Avoiding distortions due to speech coding and transmission errors in GSM ASR tasks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Mar. 1999, pp. 277–280.

[143] W. Yang and B.-Z. Yehuda, "An algorithm for detecting clipped waveforms and suggested correction procedures," *Seismological Research Letters*, vol. 81, no. 1, pp. 53–62, Jan. 2010.

[144] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "A constrained matching pursuit approach to audio declipping," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 329 –332.

[145] ——, "Audio inpainting," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

[146] S. Miura, H. Nakajima, S. Miyabe, S. Makino, T. Yamada, and K. Nakadai, "Restoration of clipped audio signal using recursive vector projection," in *IEEE Region 10 Conf. (TENCON)*, Nov. 2011, pp. 394–397.

[147] A. J. Weinstein and M. B. Wakin, "Recovering a clipped signal in sparseland," *Computing Research Repository (CoRR)*, 2011. [Online]. Available: http://arxiv.org/abs/1110.5063

[148] L. Atlas and C. P. Clark, "Clipped-waveform repair in acoustic signals using generalized linear prediction," USA Patent US8 126 578, Feb. 28, 2012.

[149] T. Otani, M. Tanaka, Y. Ota, and S. Ito, "Clipping detection device and method," USA Patent 20 100 030 555, Feb. 4, 2010.

[150] T. E. Riemer, M. S. Weiss, and M. W. Losh, "Discrete clipping detection by use of a signal matched exponentially weighted differentiator," in *Proc. IEEE Southeastcon.*, vol. 1, New Orleans, LA, USA, Apr. 1990, pp. 245–248.

[151] S. J. Godsill, P. J. Wolfe, and W. N. Fong, "Statistical model-based approaches to audio restoration and analysis," *J. of New Music Research*, vol. 30, no. 4, pp. 323–338, Nov. 2001.

[152] M. Hayakawa, T. Fukumori, M. Nakayama, and T. Nishiura, "Suppression of clipping noise in observed speech based on spectral compensation with gaussian mixture models and reference of clean speech," *J. Acoust. Soc. Am.*, vol. 133, p. 3246, 2013.

[153] J. S. Abel and J. O. Smith, III, "Restoring a clipped signal," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 1991, pp. 1745–1748.

[154] A. Dahimene, M. Noureddine, and A. Azrar, "A simple algorithm for the restoration of clipped speech signal," *Informatica*, vol. 32, pp. 183–188, 2008.

[155] S. Kitic, L. Jacques, N. Madhu, M. Hopwood, A. Spriet, and C. De Vleeschouwer, "Consistent iterative hard thresholding for signal declipping," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 5939–5943.

[156] H. Ochiai and H. Imai, "Performance analysis of deliberately clipped OFDM signals," *IEEE Trans. Commun.*, vol. 50, no. 1, pp. 89–101, Jan 2002.

[157] J. F. Kenney and E. S. Keeping, *Mathematics of Statistics, Pt. 1*, 3rd ed. Van Nostrand, Princeton, NJ, 1962, ch. 15, pp. 252–285.

[158] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hayerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. E. Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman, and C. Ludvigsen, "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994.

[159] C. F. V. Loan, *Introduction to Scientific Computing*. Prentice Hall, 2000.

[160] E. H. Moore, "On the reciprocal of the general algebraic matrix," *Bulletin of the American Mathematical Society*, vol. 26, no. 9, pp. 394–395, 1920.

[161] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Lett.*, vol. 27, pp. 861–874, 2006.

[162] D. P. Doane, "Aesthetic frequency classifications," *The American Statistician*, vol. 30, no. 4, pp. 181–183, 1976.

[163] X. S. Lin, A. W. H. Khong, and P. A. Naylor, "A forced spectral diversity algorithm for speech dereverberation in the presence of near-common zeros," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 888–899, Mar. 2012.

[164] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc. (AES)*, vol. 11, pp. 867–878, 2002.

[165] F. Xiong, S. Goetze, and B. T. Meyer, "Blind estimation of reverberation time based on spectro-temporal modulation filtering," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[166] C. Schuldt and P. Handel, "Blind low-complexity estimation of reverberation time," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct 2013, pp. 1–4.

[167] Y. Hioka, K. Furuya, K. Niwa, and Y. Haneda, "Estimation of direct-to-reverberation energy ratio based on isotropic and homogeneous propagation model," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept 2012, pp. 1–4.

[168] Audience, "Audience es800 series," Audience, Product brief, 2014. [Online]. Available: http://www.audience.com/images/AUD_eS800_ProdBrief_110414.pdf

[169] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The Reverb Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

[170] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. Intl. Conf. on Lang. Resources and Evaluation (LREC)*, N. Calzolari (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 2629–2634.

[171] Brüel and Kjær, "4190-L-001 0.5 inch free-field microphone with type 2669-L preamplifier, 3 Hz to 20 kHz, 200 V polarization," Skodsborgvej 307, DK-2850 Nærum, Denmark, 2012. [Online]. Available: http://www.bksv.com/Products/transducers/acoustic/microphones/microphone-preamplifier-combinations/4190-L-1

[172] EBU, *Loudness recommendation*, European Broadcasting Union (EBU) Recommendation R128-2014, 2014.

[173] DPA Microphones, "d:screet 4060 Omnidirectional Microphone, high-sensitivity," Gydevang 42-44 DK-3450 Allerød, Denmark, 2008. [Online]. Available: http://www.dpamicrophones.com/en/products.aspx?c=Item&category=128&item=24035

[174] M. H. Acoustics, "EM32 Eigenmike microphone array release notes (v17.0)," 25 Summit Ave, Summit, NJ 07901, USA, Oct. 2013. [Online]. Available: http://www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf

[175] Audacity, "Audacity - free multi-track audio editor and recorder," May 2000. [Online]. Available: http://audacityteam.org/

[176] P. P. Parada, D. Sharma, T. van Waterschoot, and P. A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenge," in *Proc. ACE*

Challenge Workshop, a satellite event of IEEE-WASPAA, New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1510.04616

[177] T. de M. Prego, A. A. de Lima, R. Zambrano-López, and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

[178] H. W. Löllmann, A. Brendel, P. Vary, and W. Kellermann, "Single-channel maximum-likelihood T60 estimation exploiting subband information," in *Proc. ACE Challenge Workshop, a satellite event of IEEE-WASPAA*, New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1511.04063

[179] F. Xiong, S. Goetze, and B. T. Meyer, "Joint estimation of reverberation time and direct-to-reverberation ratio from speech using auditory inspired features," in *Proc. ACE Challenge Workshop, a satellite event of IEEE-WASPAA*, New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1510.04620

[180] M. Senoussaoui, J. F. Santos, and T. H. Falk, "SRMR variants for improved blind room acoustics characterization," in *Proc. ACE Challenge Workshop, a satellite event of IEEE-WASPAA*, New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1510.04707

[181] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2014, pp. 55–59.

[182] Y. Hioka and K. Niwa, "PSD estimation in beamspace for estimating direct-to-reverberant ratio from a reverberant speech signal," in *Proc. ACE Challenge Workshop, a satellite event of IEEE-WASPAA*, New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1510.08963

[183] H. Chen, P. N. Samarasinghe, T. D. Abhayapala, and W. Zhang, "Estimation of the direct-to-reverberant energy ratio using a spherical microphone array," in *Proc. ACE Challenge Workshop, a satellite event of IEEE-WASPAA*, New Paltz, NY, USA, 2015. [Online]. Available: http://arxiv.org/abs/1510.08950

[184] F. Xiong, N. Moritz, R. Rehr, J. Anemuller, B. Meyer, T. G. G. Doclo, and S. Goetze, "Robust ASR in reverberant environments using temporal cepstrum smoothing for speech enhancement and an amplitude modulation filterbank for feature extraction," in *Proc. REVERB Challenge Workshop*, vol. 1, Florence, Italy, 2014.

[185] S. M. Ban and H. S. Kim, "Weight-space viterbi decoding based spectral subtraction for reverberant speech recognition," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1424–1428, Sept 2015.

[186] P. P. Parada, D. Sharma, P. A. Naylor, and T. v. Waterschoot, "Reverberant speech recognition exploiting clarity index estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015. [Online]. Available: http://dx.doi.org/10.1186/s13634-015-0237-7

Index