NON-INTRUSIVE QUALITY EVALUATION OF SPEECH PROCESSED IN NOISY AND REVERBERANT ENVIRONMENTS

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von
Herrn Benjamin H. R. Cauchi
geboren am 8. Dezember 1984
in Vernon (Frankreich)

Benjamin H. R. Cauchi: *Non-intrusive Quality Evaluation of Speech Processed in Noisy and Reverberant Environments*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo, *University of Oldenburg, Germany*

WEITERE GUTACHTER:
Dr.-Ing. Stefan Goetze, *University of Sheffield, United Kingdom*
Mike Brookes, *Imperial College, London, United Kingdom*

TAG DER DISPUTATION:
4. August 2021

ACKNOWLEDGMENTS

I would like to thank my supervisor Simon Doclo for his scientific advice and his precious comments during the writing process. I also want to thank him for many of the collaborations that made my PhD time so rewarding and would not have happened without his work. I would also like to thank Stefan Goetze who led the group Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media and Technology where I enjoyed working for several years. Furthermore, I would like to thank Mike Brookes for his valuable feedback when reviewing the thesis, and Steven van de Par for serving as chairman of the thesis committee.

I had the chance to collaborate with many talented researchers over the years. I want to thank Jens Schroeder, Marco Ruhland and Jan Wellmann, who were great office mates during my years at the Fraunhofer Institute. I learned a lot by exchanging with the colleagues of the Signal Processing Group of the University of Oldenburg. Among them, I would like to particularly thank Timo Gerkmann, Ina Kodrasi, Ante Jukić and Robert Rehr for their scientific input as well as for the fun we had during our various trips together.

Thanks to the DREAMS project, funded by the European Union, I worked with many researchers from Belgium, Denmark and the United Kingdom. My time in Imperial College, collaborating with DREAMS researchers and the group of Patrick Naylor, was a highlight of this project. I would as well like to thank the team of Tiago Falk in the MuSAE Lab. Working with them led to both nice trips to Montreal and fruitful collaborations. I also want to thank the OFFIS Institute, where I have been working during the last years, for giving me the freedom to finalize this thesis.

In addition, I would like to thank my parents, Paulette and Gabriel, for supporting my studies. Above all, I would like to thank my wonderful wife Katharina for her patience, for her encouragements, and for always making everything better.

Oldenburg, August 2021
Benjamin H. R. Cauchi

ABSTRACT

In many speech applications such as hands-free telephony or voice-controlled home assistants, the distance between the user and the recording microphones can be relatively large. In such a far-field scenario, the recorded microphone signals are typically corrupted by noise and reverberation, which may severely degrade the performance of speech recognition systems and reduce intelligibility and quality of speech in communication applications. In order to limit these effects, speech enhancement algorithms are typically applied. The main objective of this thesis is to develop novel speech enhancement algorithms for noisy and reverberant environments and signal-based measures to evaluate these algorithms, focusing on solutions that are applicable in realistic scenarios.

First, we propose a single-channel speech enhancement algorithm for joint noise and reverberation reduction. The proposed algorithm uses a spectral gain to enhance the input signal, where the gain is computed using a combination of a statistical room acoustics model, minimum statistics and temporal cepstrum smoothing. This single-channel spectral enhancement algorithm can be combined easily with existing beamforming techniques when multiple microphones are available. Evaluation results show that the proposed algorithm is able to improve speech recognition accuracy, when using clean as well as multi-condition training data. In addition, signal-based measures and the results of a listening test show that the proposed algorithm is beneficial in terms of both speech quality and reverberation suppression. In the REVERB Challenge, the proposed single-channel speech enhancement algorithm has obtained the best performance in terms of subjective speech quality among all submitted single-channel algorithms.

Second, we propose two non-intrusive speech quality measures that combine perceptually motivated features and predicting functions based on machine learning. The first measure uses time-averaged modulation energies as input features to a model tree. The second measure uses time-varying modulation energies as input features to a recurrent neural network in order to take the time-dependency of the test signal into account. Both measures are trained and evaluated using a dataset of perceptually evaluated signals comprising a wide range of algorithms, settings and acoustic scenarios. The results show that the speech quality measure using a recurrent neural network as predicting function outperforms existing non-intrusive measures and yields a similar performance as intrusive measures when trained and evaluated for a single category of algorithms. When trained and evaluated for several categories of algorithms, it even outperforms the intrusive benchmark measures, making it suitable for the selection of algorithms or algorithm parameters.

ZUSAMMENFASSUNG

In vielen Sprachanwendungen, wie z.B. in der Freisprech-Telefonie oder bei sprachgesteuerten Heimassistenten, kann der Abstand zwischen dem Benutzer und den Aufnahmemikrofonen relativ groß sein. In einem solchen Fernfeldszenario werden die aufgezeichneten Mikrofonsignale typischerweise durch Rauschen und Nachhall überlagert, wodurch die Leistung von Spracherkennungssystemen erheblich verringert sowie die Verständlichkeit und Qualität von Sprachsignalen in Kommunikationsanwendungen beeinträchtigt werden kann. Um diese Effekte zu verringern, werden typischerweise Sprachsignalverbesserungsalgorithmen angewendet. Hauptziel dieser Arbeit ist die Entwicklung neuartiger Sprachsignalverbesserungsalgorithmen für verrauschte und verhallte Umgebungen sowie von signalbasierten Messmethoden zur Bewertung dieser Algorithmen. Der Schwerpunkt liegt dabei auf Lösungen, die in realistischen Szenarien anwendbar sind.

Zunächst wird ein einkanaliger Sprachverbesserungsalgorithmus zur Reduzierung von Umgebungsgeräuschen und Nachhall vorgeschlagen. Der Algorithmus verwendet zur Verbesserung des Sprachsignals eine spektrale Gewichtungsfunktion, die auf Grundlage eines statistischen Raumakustikmodells, der Statistik spektraler Minima und einer zeitlichen Glättung des Cepstrums entworfen wird. In Systemen mit mehreren Mikrofonen kann der entwickelte einkanalige Algorithmus zur spektralen Sprachsignalverbesserung problemlos mit konventionellen Methoden zur räumlichen Filterung kombiniert werden, um die Leistung weiter zu verbessern. Der vorgeschlagene Algorithmus erhöht die Erkennungsgenauigkeit von Systemen zur automatischen Spracherkennung sowohl für ein Erkennertraining basierend auf ungestörten Signalen allein, wie auch für augmentierte Trainingsdaten. Ergebnisse signalbasierter Bewertungsalgorithmen, wie auch durchgeführter Hörtests mit Probanden zeigen, dass der vorgeschlagene Algorithmus sowohl hinsichtlich erreichter Sprachqualität, wie auch bei der Unterdrückung des Nachhalls von Vorteil ist. Bei der REVERB Challenge konnte der entwickelte Algorithmus die beste subjektive Sprachqualität unter allen eingereichten einkanaligen Algorithmen erreichen.

Zweitens werden in dieser Arbeit zwei Methoden zur Sprachqualitätsbewertung vorgeschlagen, die keinerlei Referenzsignale benötigen. Diese nutzen perzeptionsmotivierte Merkmale und Vorhersagefunktionen basierend auf Methoden des maschinellen Lernens. Die erste Bewertungsmethode verwendet zeitlich gemittelte Modulationsenergien als Eingangsmerkmale für ein Entscheidungsbaummodell. Die zweite Bewertungsmethode verwendet zeitlich variable Modulationsenergien als Eingangsmerkmale für ein rekurrentes neuronales Netzwerk, um die Zeitabhängigkeit des zu bewertenden Signals zu berücksichtigen. Beide Bewertungsmethoden werden anhand eines umfangreichen Datensatzes perzeptiv ausgewerteter Signale trainiert und optimiert. Der verwendete Datensatz wurde im Rahmen dieser Arbeit basierend auf einer Vielzahl von Algorithmen, Parametrisierungen und akustischen Szenarien erstellt. Die Ergebnisse zeigen, dass die vorgeschlagene Bewertungsmethode mit auf rekurrenten neuronalen Netzwerken basierender Vorhersagefunktion bestehende andere referenzlose Bewertungsmethoden übertrifft und eine ähnliche Leistung wie referenzbasierte Bewertungsmethoden liefert, wenn sie für eine einzelne Kategorie von Algorithmen trainiert und verwendet wird. Wenn sie auf Grundlage von verschiedenen Algorithmenkategorien trainiert wird, übertrifft sie sogar referenzbasierte Bewertungsmethoden und eignet sich daher für den Vergleich verschiedener Algorithmen oder Algorithmenparametern.

GLOSSARY

AI	articulation index
AME	averaged modulation energy
AMS	amplitude modulation spectrogram
ANN	artificial neural network
ANOVA	analysis of variance
ASR	automatic speech recognition
BSD	Bark spectral distortion
BSI	blind system identification
BSS	blind source separation
CART	classification and regression tree
CCR	comparison category rating
DAM	diagnostic acceptability measure
DFT	discrete Fourier transform
DMOS	degradation mean opinion score
DNN	deep neural network
DOA	direction of arrival
DRR	direct-to-reverberant ratio
FB	full batch
FCC	forced-choice comparison
FFT	fast Fourier transform
FWSSNR	frequency-weighted segmental SNR
GCC-PHAT	generalized cross correlation with phase transform
GMM	Gaussian mixture model
GSC	generalized sidelobe canceller
GWPE	generalized weighted prediction error
HMM	hidden Markov model
HTK	hidden Markov model toolkit
IIR	infinite impulse response
ISTFT	inverse short-time Fourier transform
LCMV	linearly constrained minimum variance
LLR	log-likelihood ratio
LPC	linear predictive coding
LSTM	long short-term memory

MARS	multivariate adaptive regression splines
MBSD	modified Bark spectral distortion
MCLP	multichannel linear prediction
ME	modulation energy
MFCC	Mel-frequency cepstral coefficient
MIMO	multiple input multiple output
ML	maximum likelihood
MMSE	minimum mean square error
MOS	mean opinion score
MS	minimum statistics
MUSHRA	multiple stimuli test with hidden reference and anchor
MUSIC	multiple signal classification
MVDR	minimum variance distortionless response
MWF	multichannel Wiener filter
NMF	non-negative matrix factorization
PDF	probability density function
PEMO-Q	perception model for quality
PESQ	perceptual evaluation of speech quality
PLPC	perceptual linear prediction coefficient
POLQA	perceptual objective listening quality assessment
PSD	power spectral density
PSM	perceptual similarity measure
RIR	room impulse response
RLS	recursive least squares
RMSE	root mean square error
RNN	recurrent neural network
RT	real time
SD	spectral distortion
SII	speech intelligibility index
SIR	signal-to-interference ratio
SNR	signal-to-noise ratio
SPP	speech presence probability
SRMR	speech-to-reverberation modulation energy ratio
SRMR$_{\text{norm}}$	normalized SRMR
SRT	speech reception threshold
SSI	supervised system identification
SSNR	segmental SNR
STFT	short-time Fourier transform
STI	speech transmission index
STOI	short-time objective intelligibility

UB	utterance-based
UN	unprocessed
VAD	voice activity detector
WER	word error rate
WPE	weighted prediction error

CONTENTS

1 Introduction 1
- 1.1 Speech recording in noisy and reverberant environments 2
- 1.2 Speech enhancement in the spectral domain 6
  - 1.2.1 Single-channel spectral enhancement 7
  - 1.2.2 Spatial processing and spectral enhancement 8
  - 1.2.3 Direct inverse filtering 10
- 1.3 Evaluation of speech enhancement techniques 11
  - 1.3.1 Automatic speech recognition 12
  - 1.3.2 Perceptual evaluation 14
  - 1.3.3 Signal-based measures 16
- 1.4 Outline of the thesis and main contributions 18

2 Problem formulation and considered performance measures 21
- 2.1 Problem formulation 21
- 2.2 Speech enhancement benchmark 23
  - 2.2.1 Considered ASR system 24
  - 2.2.2 Considered quality measures 25
- 2.3 Quality measures benchmark 28
- 2.4 Summary 29

3 Speech enhancement in single- and multichannel scenarios 31
- 3.1 System overview and beamformer 33
  - 3.1.1 MVDR beamforming 33
  - 3.1.2 Noise coherence matrix estimation 34
  - 3.1.3 DOA estimation 35
- 3.2 Single-channel spectral enhancement 36
  - 3.2.1 Spectral gain estimation 37
  - 3.2.2 Noise PSD estimation 38
  - 3.2.3 Speech PSD estimation 39
  - 3.2.4 Reverberant PSD estimation 41
- 3.3 Experimental setup 42
  - 3.3.1 Algorithm settings 42
  - 3.3.2 Observations on beamformer design 42
- 3.4 Results 43
  - 3.4.1 Speech recognition performance 43
  - 3.4.2 Signal-based measures 46
  - 3.4.3 Perceptual evaluation 48
- 3.5 Summary 53

4 Non-intrusive quality prediction for processed speech 55
4.1 Proposed approach . . . 57
4.1.1 Considered features . . . 57
4.1.2 Model tree as predicting function . . . 59
4.1.3 LSTM network as predicting function . . . 60
4.2 Dataset . . . 63
4.2.1 Considered algorithms . . . 64
4.2.2 Perceptual evaluation . . . 68
4.3 Experiments . . . 70
4.3.1 Benchmark and figures of merit . . . 70
4.3.2 Training framework . . . 71
4.4 Results . . . 73
4.5 Summary . . . 76
5 Conclusion and further research 79
5.1 Conclusion . . . 79
5.2 Further research directions . . . 80
A Statistical analysis 83
B Observations on the LSTM measure 85
B.1 Performance on our dataset . . . 86
B.2 Performance on REVERB challenge dataset . . . 87

BIBLIOGRAPHY 91

LIST OF FIGURES

Fig. 1.1 Spectrograms of different noise types 3
Fig. 1.2 Reflections of a single speech source in an enclosure 4
Fig. 1.3 Example of room impulse response 5
Fig. 1.4 Effect of direct path, early reflections, all reflections and noise on the spectrogram of one exemplary speech signal 6
Fig. 1.5 Typical single-channel spectral enhancement 7
Fig. 1.6 Typical combination of beamforming and single-channel spectral enhancement 9
Fig. 1.7 Typical combination of MIMO processing with beamforming and spectral enhancement 11
Fig. 1.8 Overview of a typical ASR system 13
Fig. 1.9 Example of user interface used to conduct a MUSHRA test 15
Fig. 1.10 Overview of signal-based speech quality measures 16
Fig. 1.11 Overview of the thesis. 19
Fig. 2.1 Signal model for a single speech source recorded in a noisy and reverberant environment 22
Fig. 2.2 Overview of the REVERB challenge dataset 24
Fig. 2.3 Example of sigmoidal mapping used to compute ρ_{sig} 27
Fig. 2.4 Error thresholding to compute ϵ-RMSE 28
Fig. 3.1 Overview of the proposed algorithm 33
Fig. 3.2 Overview of the proposed single-channel speech enhancement . . . 36
Fig. 3.3 Power of noise as well as of clean and reverberant speech 39
Fig. 3.4 DOA estimation error and beampattern of the MVDR beamformer 41
Fig. 3.5 Speech recognition accuracy on simulated data 43
Fig. 3.6 Speech recognition accuracy on real data 44
Fig. 3.7 Intrusive speech quality measures computed from simulated data . 45
Fig. 3.8 Non-intrusive speech quality measures computed using simulated data 46
Fig. 3.9 Non-intrusive speech quality measures computed using real data . . 47
Fig. 3.10 MUSHRA scores comparing different configurations of the proposed algorithm 49
Fig. 3.11 MUSHRA scores for single-channel algorithms obtained during the test conducted online by the REVERB challenge organizers 51
Fig. 3.12 MUSHRA scores for multichannel algorithms obtained during the test conducted online by the REVERB challenge organizers 52
Fig. 4.1 Overview of the feature extraction 58
Fig. 4.2 Modulation energies extracted from a short speech segment 59
Fig. 4.3 Overview of the updates applied by an LSTM-layer 62
Fig. 4.4 Overview of the network used as predicting function 63

Fig. 4.5 MUSHRA scores for all combinations of acoustic conditions, algorithms and settings . . . 69
Fig. 4.6 Performance of the considered measures trained for one algorithm at a time . . . 72
Fig. 4.7 Performance of the considered measures trained for all considered algorithms . . . 73
Fig. 4.8 Scatter plot of predicted quality using either SRMRnorm or the LSTM-based measure . . . 74
Fig. 4.9 Performance of the proposed measures in presence of mismatch between training and testing sets . . . 75
Fig. 4.10 Scatter plot of predicted quality using either Tree or the LSTM-based measure . . . 76
Fig. B.1 Influence of dropout and network size on obtained correlation . . . 86
Fig. B.2 Performance obtained when using the best dropout rate for each considered network size . . . 87
Fig. B.3 Correlation obtained using the single-channel part of the REVERB challenge dataset . . . 88
Fig. B.4 Correlation obtained using the multichannel part of the REVERB challenge dataset . . . 89

LIST OF TABLES

Table 2.1 Recording distance, T_{60} and D_{50} for all recordings included in the speech enhancement benchmark 25
Table 3.1 Results of the Friedman's test conducted to assess the significance of the MUSHRA scores 50
Table 4.1 Expected behavior of the considered algorithms. 64
Table 4.2 RIRs used to generate the perceptually evaluated signals 65
Table 4.3 Overview of the dataset of perceptually evaluated signals 68

1

INTRODUCTION

Speech is one of the most natural forms of communication to exchange information or to express emotions [1], making it particularly suited to design intuitive human-computer interfaces [2,3]. In addition, due to the progress in hardware and software engineering mobile computing devices, such as smartphones, tablets and laptops are nowadays omnipresent, and allow for advanced signal processing algorithms and automatic speech recognition (ASR) systems. Thanks to these developments and an ever faster and larger internet coverage, speech communication and speech-based interfaces are used by a vast number of users. Applications include hands-free telephony, teleconferencing, hearing devices and voice control, in diverse locations, e.g., at home, in public venues or at workplaces. Regardless of the encountered situation or acoustic environment, users expect a satisfactory speech communication or ASR performance.

In so-called far-field scenarios, the distance between the user and the recording microphone(s) is relatively large. The far-field scenario is highly relevant for many common speech applications such as hands-free phone calls, voice control of home assistants or communication using hearing devices. While applications that can be used in far-field scenarios are convenient for the user, these scenarios present greater challenges than close-talk scenarios, where the microphones are close to the mouth of the target speaker. Indeed, in far-field scenarios the microphones do not only record the desired speech of the target speaker but also undesired noise generated by other sound sources, e.g., concurrent speakers or music, which may greatly degrade speech quality or ASR performance [4–6]. In addition, in an enclosed environment the recorded speech signal is reverberant, i.e., it contains not only the direct speech of the target speaker but also reflections against walls, ceiling and objects in the room [7]. Although it has been shown that early reflections of the target speech can be beneficial [8–11], large levels of reverberation can degrade speech quality and intelligibility, as well as severely degrade the performance of ASR systems [4–6].

Hence, far-field speech applications require speech enhancement algorithms that provide effective noise reduction and dereverberation while preserving the target speech and, possibly, its early reflections [12]. In the last decades, several single- and multi-channel noise reduction algorithms have been proposed [13–23], e.g., assuming that the undesired noise is uncorrelated with the target speech and/or assuming that the spatial characteristics of the noise are different from the spatial characteristics of the target speaker. Since reverberation is highly correlated with the speech to be preserved, dereverberation arguably presents a greater challenge than noise reduction. Nevertheless, a variety of single- and multichannel speech dereverberation algorithms have been proposed, e.g., assuming certain temporal or spatial characteristics for reverberation [24–33]. Obviously, the ultimate goal is to design speech enhancement algorithms that achieve joint noise and reverberation reduction.

Designing speech enhancement algorithms and selecting the best algorithm for a given application requires reliable performance measures. For ASR applications, the performance of a speech enhancement algorithm is typically measured as the effect on the word recognition accuracy [34, 35]. For speech communication applications, the performance of a speech enhancement algorithm is typically measured as the effect on speech quality and/or speech intelligibility. Subjective listening tests are usually considered the most reliable way to measure these attributes, but they can be costly and time-consuming and are not always applicable [36, 37]. Alternatively, signal-based measures for speech quality and speech intelligibility have been proposed. These measures can be either intrusive, requiring a (clean) reference signal, or non-intrusive, which can be computed using only the test-signal [38–40]. A reliable non-intrusive signal-based measure that could be used to compare a wide range of speech enhancement algorithms would be a great asset.

This thesis focuses on speech enhancement algorithms aiming at improving the quality of speech recorded in the presence of both noise and reverberation. The **first objective** is to develop novel single- and multichannel algorithms for joint noise and reverberation reduction. The **second objective** is to develop non-intrusive signal-based measures of speech quality that can be used to reliably evaluate speech enhancement algorithms designed to operate in noisy and reverberant environments. The remainder of this chapter is organized as follows. In Section 1.1 we describe the acoustic scenarios considered in this thesis. In Section 1.2 we provide an overview of single- and multichannel speech enhancement algorithms in the spectral domain. In Section 1.3 we discuss existing performance measures to evaluate speech enhancement algorithms. In Section 1.4 we present an overview of the thesis and its main contributions.

1.1 Speech recording in noisy and reverberant environments

The work presented in this thesis considers scenarios in which the speech of a single target speaker is recorded in a reverberant room in the presence of noise. Noise is composed of sounds whose sources are anything but the target speaker and can hence be assumed to be uncorrelated with the speech signal to be preserved. As noise sources can be extremely diverse, noise suppression algorithms are typically designed to be applicable to a wide range of unknown noise types. Commonly considered noise types are, e.g., babble noise, corresponding to multiple speakers talking simultaneously, or fan noise, chosen for its slowly varying spectro-temporal characteristics. In noise suppression algorithms, it is often assumed that the noise signal to be suppressed is more stationary than the target speech to be preserved. When stationarity of the noise cannot be assumed, other properties have to be taken into account. In some specific applications, noise suppression can be designed to

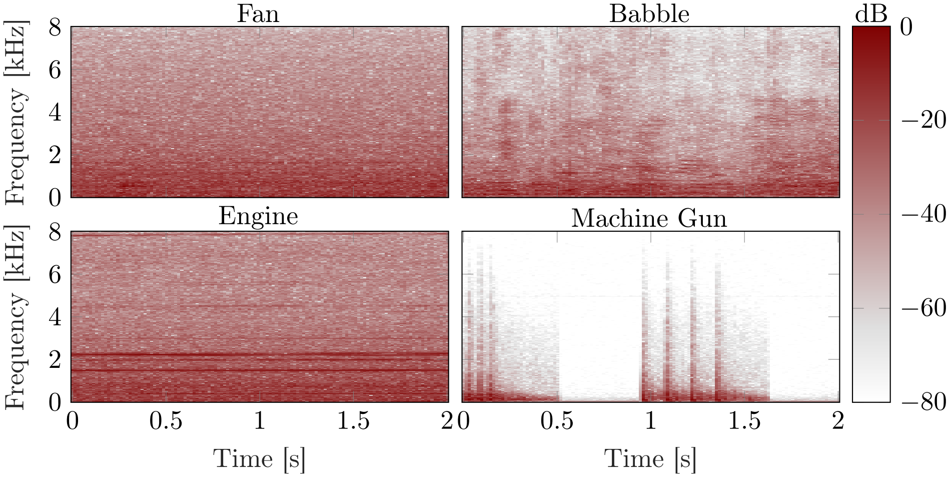

Fig. 1.1: Spectrograms of different noise types. The spectral characteristics of stationary noises such as fan noise (top left) or engine noise (bottom left) are typically easier to estimate than the spectral characteristics of fast varying noises such as machine gun noise (bottom right) or babble noise (top right).

suppress a particular type of noise, e.g., ego-noise generated by moving parts of a robot [41–43], or keyboard strokes [44–46]. However, such approaches are limited to a narrow range of scenarios. The stark difference between different noise types can be illustrated by their time-frequency representations (spectrograms), depicted in Fig. 1.1. When multiple microphones are available, one can exploit the spatial characteristics of the noise in addition to the spectral characteristics, e.g., by designing a noise suppression algorithm that is optimized for a certain (assumed) noise field. Examples of commonly assumed noise fields are homogeneous, incoherent or coherent noise fields. In an homogeneous noise field the power spectral densities (power spectral densitys (PSDs)) recorded in all microphones are equal, which can be a good model for babble noise. In an incoherent noise field, noise signals recorded in different microphones are uncorrelated, which can be a good model for microphone self-noise. On the contrary, a coherent noise field can be a good model for the speech of an interfering speaker [47].

Reverberation consists of a superposition of reflections of the speech source against the surfaces and objects in the room. These reflections can be classified into three categories, namely direct speech, early reflections and late reflections, as illustrated in Fig. 1.2. The direct speech signal is equal to the source signal, up to an attenuation and a propagation delay. Reflections recorded within a few tens of milliseconds are typically considered early reflections. Although early reflections may introduce spectral coloration, they are often considered beneficial for speech intelligibility. Indeed, within such a short time window, the direct speech signal and early reflections are perceived as a single signal whose energy is larger than the energy of the direct speech alone. In noisy environments, this results in a larger perceived signal-to-noise ratio (SNR) [8–11, 48]. In addition, the early reflections contain spatial information

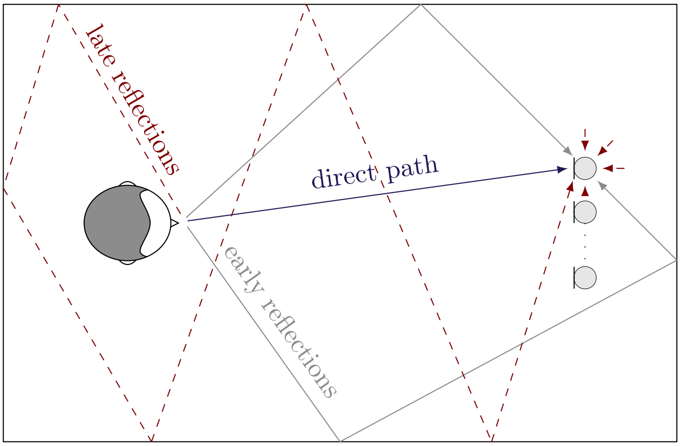

Fig. 1.2: In a room with a single speech source, the signal recorded by each microphone of an array contains the direct speech as well as early and late reflections.

that can be used to localize the speech source of interest [49–52] or apply spatial filtering [50–53]. Reflections recorded more than 50 to 80 milliseconds after the source are typically considered late reflections. It is often assumed that the spatial characteristics of these reflections is isotropic/diffuse, i.e., that their direction of arrival (DOA) is uniformly distributed [7, 24]. Late reflections may have a severe detrimental effect on the performance of ASR systems [6, 31] as well as on the perceived speech quality and intelligibility in speech communication applications [4, 5, 11, 54, 55]. The impact on speech intelligibility can be particularly noticeable for people suffering from hearing loss and for non-native speakers [5, 56–58]. Consequently, the main objective of dereverberation algorithms is to suppress late reflections. Contrary to noise, these reflections can be highly correlated with the source of interest and suppressing them while preserving the target speech (and early reflections) can be challenging [26].

The acoustical propagation between a source and a receiver, i.e., a microphone, can be characterised by the so-called room impulse response (RIR), which is exemplary depicted in Fig. 1.3. While the direct path represents the delay and attenuation of the source signal, the early reflections consist of numerous but defined impulses and the late reflections consist of densely-spaced impulses, often modelled as a random process whose amplitude is exponentially decaying [24]. Various approaches have been proposed for RIR estimation, where one needs to distinguish between supervised system identification (SSI), requiring the source signal to be known, and blind system identification (BSI). SSI techniques vary in terms of the used source signal and the method used to extract the RIR from the recorded microphone signal. Common examples include maximum length sequences [59,60], inverse repeated sequences [61], time-stretched pulses [62] or sine-sweeps [63, 64]. However, the exact RIR can vary greatly with the location of the source and the microphone, or even with changes in temperature [7, 65]. BSI techniques aim at estimating RIRs from reverberant and potentially noisy microphone signals without knowledge of the

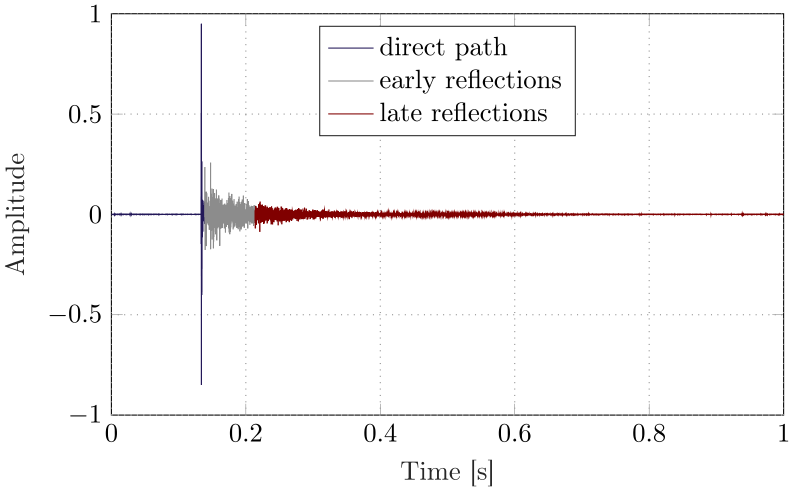

Fig. 1.3: The RIR contains information about the acoustics of a given room.

source signal. BSI approaches are usually based on higher-order statistics [66] or on second-order statistics [67–70] but, despite large efforts, the obtained RIR estimates can be highly inaccurate due to the noise conditions or numerical instabilities.

Some dereverberation algorithms make use of RIR estimates obtained using BSI to enhance the recorded signal. However, attempts at improving robustness of these so-called multichannel equalization algorithms against RIR estimation error [71–74] have not been successful enough for such algorithms to be applicable in realistic scenarios.

In practice, the design of dereverberation algorithms relies on RIR models, such as the infinite impulse response (IIR) model, the common acoustical poles and zeros model or orthonormal basis functions [75–77]. In addition, algorithms can take advantage of a few parameters used to describe reverberation. Among those, the most commonly found in the literature are the reverberation time, the direct-to-reverberant ratio (DRR) and the clarity [26]. The reverberation time represents the amount of reverberation in a room and depends only on the room characteristics, i.e., is independent of source and microphone positions. The reverberation time is typically defined as T_{60}, i.e., the time for the reverberant energy to decay by 60 dB after deactivation of the sound source. The DRR is defined as the ratio between the energy of the direct speech and the energy of the reverberant speech and, in addition to the room characteristics, depends on the source and microphone positions. The clarity is similar to the DRR, but considers both direct and early reverberant speech as the signal of interest. The clarity is often measures as D_{50}, i.e., the ratio between the energy of the first 50 ms of the RIR and the energy of the complete RIR. Reverberation time, DRR and clarity can be computed from a RIR [78] but can also be estimated from noisy reverberant signals [79, 80], making them useful parameters for dereverberation algorithms. Some dereverberation algorithms have been designed by modelling the time-domain RIR as a stationary noise multiplied by

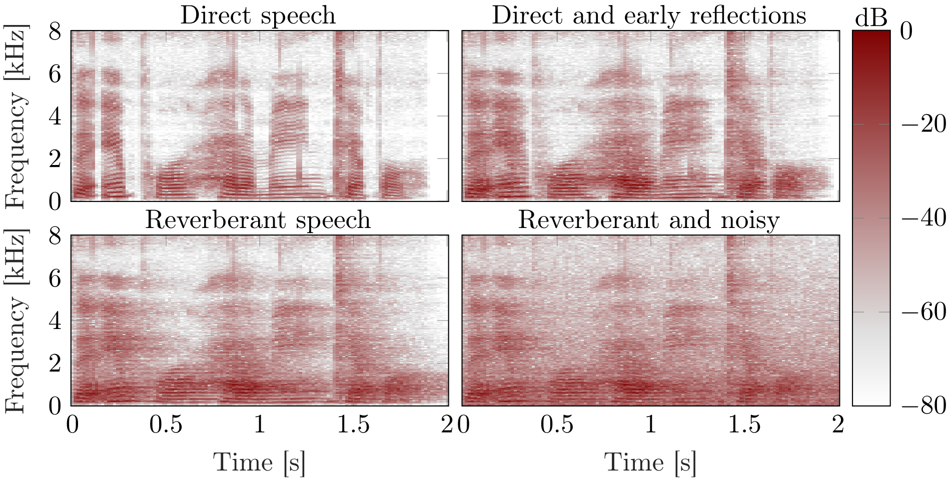

Fig. 1.4: Effect of direct path, early reflections, all reflections and noise on the spectrogram of one exemplary speech signal. The combination of both noise and reverberation can mask the target speech.

an exponential decay whose rate depends on the T_{60} [24], and have been extended to take into account the DRR [29, 81]. When multiple microphones are available, models can be based on assumed spatial properties of reverberation [33, 82].

1.2 Speech enhancement in the spectral domain

This thesis considers the design and evaluation of speech enhancement algorithms that estimate the signal of interest in the spectral domain, making use of the noise and reverberation models discussed in the previous section.

Speech enhancement in the spectral domain relies on transforming the input signal from the time domain to the spectral domain, i.e., to a time-frequency representation. Depending on the chosen transform, the time-domain signal may or may not be reconstructible from its time-frequency representation. Such reconstruction is not required for ASR applications, but obviously necessary for speech communication applications. In this thesis, the discrete input signals are transformed to the spectral domain using the short-time Fourier transform (STFT), one of the most commonly used time-frequency transforms for speech enhancement [23, 83]. First, the discrete-time input signal is divided into overlapping segments, with a typical length of 10 to 30 ms. Then, a so-called analysis window is applied before computing the frequency representation of this windowed segment with the Fourier transform, typically using the fast Fourier transform (FFT) algorithm. The complex-valued STFT representation can be transformed back to the time domain using the inverse short-time Fourier transform (ISTFT), i.e., by applying an inverse FFT and a so-called synthesis window. The STFT and its inverse do not introduce distortion provided that the overlap factor as well as the analysis and synthesis windows are

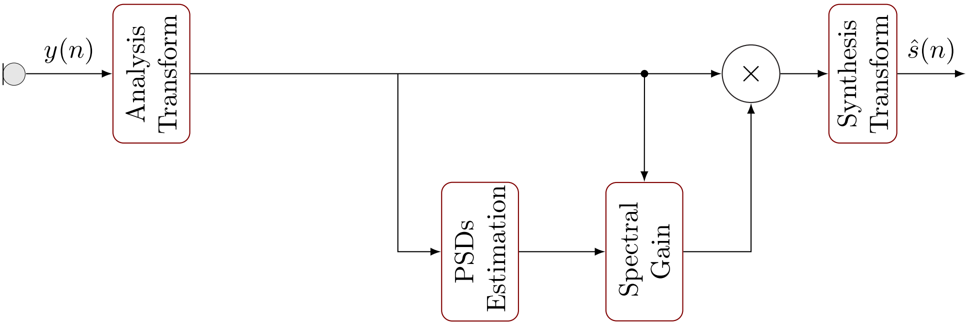

Fig. 1.5: In a typical single-channel spectral enhancement algorithm, a spectral gain is applied to the time-frequency representation of the microphone signal $y(n)$ to obtain the enhanced signal $\hat{s}(n)$. The spectral gain is typically computed based on the estimated PSDs of the interference to be suppressed and the speech to be preserved.

chosen appropriately [84]. For an exemplary speech signal and RIR, Fig. 1.4 depicts the spectrograms, i.e., the squared amplitude of the STFT coefficients, for direct speech, direct speech and early reflections, reverberant speech and noisy reverberant speech. The algorithms described in the following sections aim at estimating the direct speech STFT coefficients, with or without early reflections, from the noisy reverberant STFT coefficients in all available microphones.

The remainder of this section is organized as follows. In Subsection 1.2.1 we describe single-channel spectral enhancement with a gain, that can be applied when only one microphone is available. In Subsection 1.2.2 we discuss how, when multiple microphones are available, speech enhancement can exploit spectral information as well as spatio-temporal properties. In Subsection 1.2.3 we introduce direct inverse filtering that, typically applied before spatial filtering, can further improve dereverberation.

1.2.1 *Single-channel spectral enhancement*

Many single-channel spectral enhancement approaches have been developed aiming at noise suppression [15,20,85,86] or dereverberation [25,29,87]. Fig. 1.5 depicts the block diagram of a typical spectral enhancement algorithm, where a (real-valued) spectral gain is applied to the time-frequency representation of the microphone signal. The computation of the spectral gain typically requires an estimate of the PSDs of the interference to be suppressed, i.e., noise and/or reverberation. After applying this gain, an inverse transform is used to synthesise the enhanced signal.

To estimate the noise PSD, several approaches have been proposed. A simple approach estimates the noise PSD during segments where speech is absent. This approach requires a voice activity detector (VAD) [88,89] and performs poorly for nonstationary noise. Improved noise PSD estimators were proposed in [90–92] based on minimum statistics (MS), i.e., assuming that in each time-frequency bin the noise power can be estimated as the minimum input power within a sliding, compensated by a bias factor. Although MS-based approaches remove the need for a VAD, they still struggle to track the PSD of rapidly varying noise. More recent noise PSD estimators, e.g., based on frequency-dependent speech presence probability (SPP), have focused on improving the tracking performance for non-stationary noise [93, 94].

When applying spectral enhancement for dereverberation, the late reverberant PSD needs to be estimated instead of the noise PSD. In [87] it has been proposed to estimate the late reverberant PSD [33] based on the temporal exponential decay model of RIRs proposed in [24]. This estimator has been updated in [29,81] to avoid overestimation of the late reverberant PSD by modelling the early reflections as a random process independent of the late reflections. Although probabilistic models have been used to jointly estimate noise and reverberation PSDs [95, 96], these estimators have a larger complexity and a similar performance as approaches that estimate the noise PSD and the late reverberant PSD separately. In the later case, noise and reverberation are assumed to be uncorrelated and the sum of the noise and late reverberant PSDs is used as the PSD of the interference to be suppressed.

Based on the estimated PSD of the interference, different spectral gain have been proposed. The more straightforward, but still widely used, spectral gain is the so-called Wiener gain [15, 20, 83]. Aiming at improving the denoising and dereverberation performance while reducing the amount of artefacts, other spectral gain functions have been derived, typically using different statistical models of the speech and interference [20, 97–99].

It should be noted that many recent approaches have explored the use of machine learning to improve the performance of spectral enhancement. For example, non-negative matrix factorization (NMF) has been used for denoising [100] as well as for dereverberation [101]. Many recent approaches make use of deep neural networks (DNNs) and show promising performance [102]. Such DNN-based approaches can be designed to estimate the spectral coefficients of the clean signal [103, 104] or, more similarly to the previously described approaches, to estimate a spectral gain [105–107].

Despite these advances, single-channel spectral enhancement has its drawbacks. The suppression of noise and reverberation of single-channel approaches can often introduce artefacts, such as so-called musical noise. Consequently, the maximum level of suppression is, in practice, limited in order to obtain a trade-off between interference suppression and target speech preservation. In addition, single-channel spectral enhancement often only enhances the amplitude of the spectral coefficients and the output signal is synthesised using the input (noisy and reverberant) phase. This may significantly limit the performance of single-channel spectral enhancement, particularly for dereverberation. Consequently, spectral enhancement is, when possible, often combined with spatial processing.

1.2.2 *Spatial processing and spectral enhancement*

When multiple microphones are available, spatial filtering can be combined with spectral enhancement in order to exploit both spatial information and spectro-

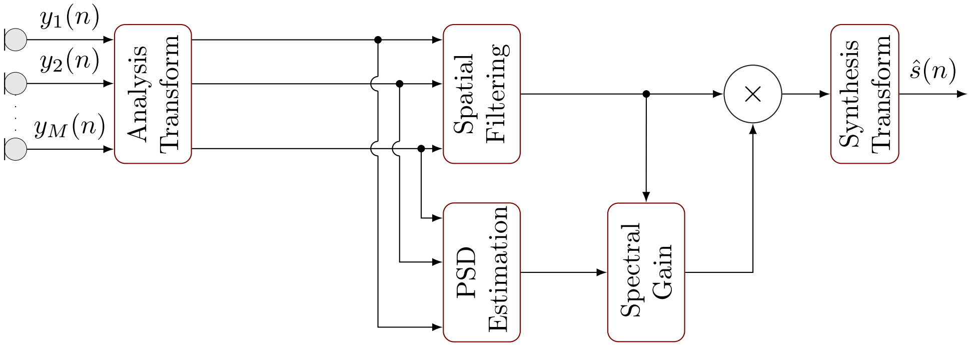

Fig. 1.6: When multiple microphones are available, speech enhancement algorithms can exploit the spatial properties of the signals recorded using M microphones. Spatial filtering is often combined with single-channel enhancement, for which required PSD estimates may be computed from the multichannel input.

temporal properties. Beamforming is a type of spatial filtering that is known to improve the performance of ASR systems and of speech communication. Beamforming has been widely used for noise reduction [16, 21, 22, 108–112], aiming at preserving the target speech source from a given DOA while suppressing interfering sources and noise from other directions. To achieve this, the multichannel input signals are typically linearly filtered before being summed together in a, so-called, filter-and-sum structure. This can be done either in the time domain or, as in this thesis, in the time-frequency domain.

Several signal-independent and signal-dependent beamformers have been proposed The simplest signal-independent beamformer is the delay-and-sum beamformer, which delays the input signals in order to time-align the target speech and where the output signal is equal to the mean of these delayed signals. The delay-and-sum beamformer optimally suppresses incoherent interference, i.e., interference signals that are uncorrelated between the microphones [109]. Another frequently used signal-independent beamformer is the superdirective beamformer [108, 113], which is optimal for an isotropic noise field. However, since in practice the noise field is never perfectly uncorrelated or isotropic and may change over time, using the estimated spatial properties of the interference in signal-dependent beamformers may improve the amount of interference suppression. Widely used signal-dependent beamformers are, e.g., the linearly constrained minimum variance (LCMV) beamformer [114] or the minimum variance distortionless response (MVDR) beamformer [108,113,115]. Since the DOA of the target speech source is usually unknown, it needs to be estimated. Although source localization, e.g., based on generalized cross correlation with phase transform (GCC-PHAT) [116] or multiple signal classification (MUSIC) [117, 118] typically performs well, DOA estimation errors may occur in challenging noise conditions, resulting in a beamformer potentially suppressing the signal of interest.

Typically, the performance of beamformers is rather low when using few microphones or small microphone arrays or when the interference is diffuse, as is the case with reverberation [119]. Consequently, spectral enhancement is often applied to the output of the beamformer to suppress residual interference. Such a combination can be interpreted as the well-known multichannel Wiener filter (MWF) [120, 121]. The PSDs required to compute the spectral gain can either be computed from the beamformer output, similarly as in the single-channel case discussed in Subsection 1.2.1, or from the multichannel input as depicted in Fig. 1.6. In challenging acoustic scenarios with fast varying noise, noise PSD estimators taking advantage of the multichannel input such as in [122, 123] can be beneficial. Several multichannel estimators have also been proposed to estimate the late reverberant PSD. In [25], it has been proposed to estimate the late reverberant PSD as the average of the estimates obtained in each channel using the single-channel estimator from [29, 81]. Though such estimate is only valid if the source-to-microphone distance is roughly equal for all microphones, this approach can generally be used for small-sized microphone arrays. In other approaches, it is assumed that late reverberation can be modelled as a diffuse sound field with a time-invariant spatial coherence matrix, and several methods have been proposed to estimate the time-varying diffuse PSD, e.g., based on maximum likelihood estimation [124], Frobenius norm minimization [125] or eigenvalue decomposition [126].

The progress made in the field of machine learning has motivated the development of beamforming methods based on deep learning [127, 128]. These DNN-based approaches are now widely used in ASR applications in which both spatial and spectral filtering are typically trained jointly with the network used for ASR. Recent ASR challenges have showcased their applicability and efficacy for those applications [34, 35].

Although combinations of beamforming and spectral enhancement typically result in a good noise and interference suppression, their performance in terms of reverberation suppression may still be limited. For this purpose, such algorithms are increasingly combined with a dereverberation stage, e.g., based on multichannel linear prediction.

1.2.3 *Direct inverse filtering*

In order to improve dereverberation performance of multichannel speech enhancement systems, a multiple input multiple output (MIMO) algorithm is often applied prior to spatial filtering, as depicted in Fig. 1.7. In this case dereverberation can be achieved, e.g, by estimating the reverberant component in each microphone signal and subtracting it from the microphone signals. This reverberant component is often estimated using multichannel linear prediction (MCLP) [27, 32, 129–137].

Contrary to approaches such as multichannel equalization [71, 72], direct inverse filtering approaches such as MCLP do not use knowledge of the RIRs between the source and the microphones, such that designing the required filters can be particularly challenging and typically relies on some assumed characteristics of the desired speech signal [27, 138]. Early MCLP methods suffered from two main limitations that hindered their application to realistic scenarios. First, the filters to estimate the reverberant component to be suppressed aimed at minimizing the energy of

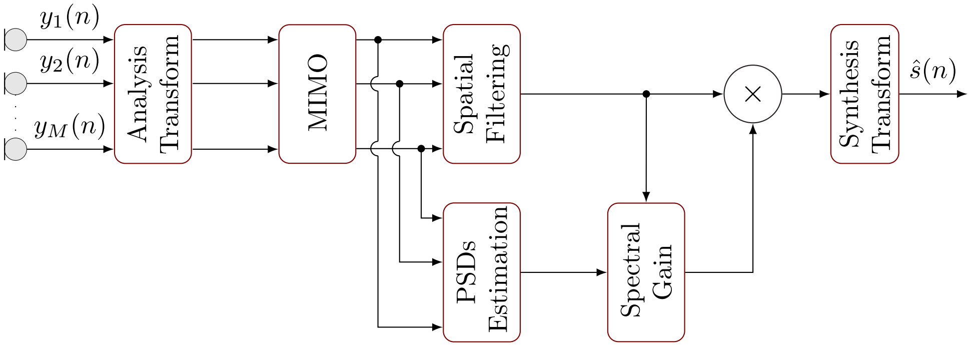

Fig. 1.7: A MIMO algorithm, such as multichannel linear prediction, is often applied to the multichannel signals prior to spatial filtering. A single-channel spectral enhancement algorithm is typically used to suppress the residual interference.

the output signal based on the assumption that the speech to be preserved was temporally white [129–131], which could lead to over-estimation of the reverberant component and consequently to signal/speech distortion. Second, the computation of these filters required the complete utterance to be processed, i.e., so-called batch processing, making it unusable in real-time applications.

Many extensions of MCLP methods have been proposed to address these drawbacks. First, the over-estimation of the reverberant component can be averted by using a prediction delay [131] and by using weighted prediction error (WPE) to take into account the time-varying nature of speech signals [132]. Second, in order to overcome the need for batch processing, it has been proposed to estimate the filters online using the recursive least squares (RLS) algorithm or Kalman filtering [133,135]. However, these online approches could still overestimate the reverberant component in the presence of fast changing RIRs, e.g., if the target speech source is moving. This has been addressed in [136] by constraining the power of the estimated reverberant component based on an estimate of the late reverberant PSD (see Subsection 1.2.1). Recent publications have focused on improving the robustness against a change of speaker position, e.g., by combining the RLS algorithm with an adaptive smoothing factor [137]. Finally, several applications of MCLP have been proposed to jointly address dereverberation and noise suppression [139–141]. Combinations of MCLP, spatial filtering and spectral enhancement have been shown to be particularly effective for ASR and speech enhancement, as illustrated by the results of international challenges [31, 35] and their application to consumer products [142, 143].

1.3 Evaluation of speech enhancement techniques

Depending on the specific application, the speech enhancement algorithms presented in Section 1.2 may need to be modified and their parameters need to be tuned. Practical considerations, such as the number of available microphones or computational constraints, often dictate many of the choices to be made. In addition, the choice of the optimal speech enhancement algorithm and of its parameters may greatly depend on the acoustic environment in which the algorithm will operate. For example, while most speech communication applications aim at improving the perceived speech quality, processing in hearing aids typically focuses on increasing speech intelligibility and preprocessing for ASR applications aims at improving the accuracy of a specific ASR system.

The remainder of this section provides an overview of the evaluation of speech enhancement algorithms for such applications and is organized as follows. In Subsection 1.3.1, we present an overview of a typical ASR system. In Subsection 1.3.2, we describe the use of listening tests to measure speech intelligibility or speech quality. In Subsection 1.3.3, we introduce signal-based measures of speech intelligibility and speech quality.

1.3.1 *Automatic speech recognition*

ASR aims at recognising a speech utterance from a recorded audio signal and can provide an intuitive, hands-free interface to control a multitude of devices. Although the performance of ASR systems was previously too poor to be implemented in consumer products, this performance has rapidly improved during the past two decades, such that ASR systems have now become ubiquitous [144]. For example, ASR is used for convenience in mobile phones, tablets and computers, as well as in more critical applications, e.g., to assist healthcare workers in protocoling their tasks [145], to control surgical robots [146] or to support speech therapy [147]. In order for ASR systems to be accepted by the end users, they need to perform satisfactorily in noisy and reverberant environments or when the target speaker is far from the microphones. ASR systems yielding such performance rely on state-of-the-art methods from several disciplines, among which speech enhancement is used to increase the ASR robustness in adverse acoustic environments.

An overview of a typical ASR system is depicted in Fig. 1.8. In such a system, the speech from the user is recorded using one or multiple microphones and speech enhancement is performed to suppress noise and reverberation that could be detrimental to the performance of the ASR system. Even when multiple microphones are used, the output of the speech enhancement step is often a single channel signal. This enhanced signal is used as the input to the feature extraction stage, sometimes referred to as the ASR front-end, which extracts features that are relevant to the recognition task. Features used in ASR systems are chosen to represent the useful information present in the input signal while being invariant to irrelevant factors, such as pitch, speaker, processing artefacts and interferences that have not been suppressed by the speech enhancement step. Commonly used ASR features are the Mel-frequency cepstral coefficients (MFCCs) [148], perceptual linear prediction coefficients (PLPCs) [149] or amplitude modulation spectrograms (AMSs) features [150]. Generally, the impact that the choice of feature has on the performance of the overall system decreases when a more advanced decoding stage is used.

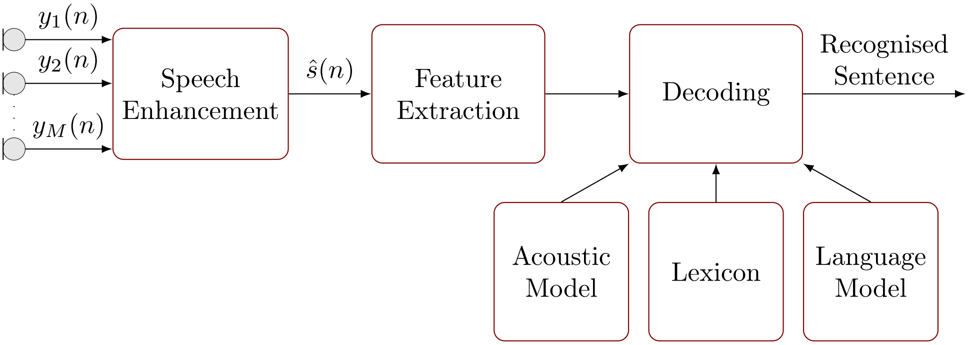

Fig. 1.8: Overview of a typical ASR system. The single- or multichannel signal is processed by a speech enhancement algorithm and features are extracted from its output. The decoding stage recognises a sentence from these features based on an acoustic model, a lexicon and a language model.

The decoding stage can be roughly described by three components [144]. First, the acoustic model represents the likelihood of a given feature sequence given a sequence of phonemes, i.e., short segments of speech. Second, the lexicon represents the complete set of words, i.e., sequences of phonemes, that can be recognised. Third, the language model represents the probability that a given sentence, i.e., sequence of words, is present in the considered language. By combining these three components, the probability that a given sentence is present in the speech signal can be computed from the input features, typically using Bayesian statistics.

Early ASR systems used hidden Markov models (HMMs) for the acoustic model and would often rely on manually defined lexicon and language models. State-of-the-art ASR systems, as implemented in many end user devices, nowadays largely rely on DNNs. Training such DNNs typically requires large amounts of labeled speech data to train the acoustic model while lexicon and language models are often trained on text corpora. The accuracy of ASR systems can be improved by using model adaptation to improve robustness in adverse acoustic environments [144] or by extracting features from multichannel signals [151]. It is worth noting that more recent end-to-end approaches use a single DNN instead of separately trained acoustic and language models [152, 153].

When comparing speech enhancement algorithms based on ASR performance, one can obviously only compare scores obtained on the same test set with the same combination of features and decoding. In addition, one should keep in mind that the impact of speech enhancement on ASR performance decreases when a more advanced decoding stage is used. However, as long as such considerations are kept in mind, the performance of speech enhancement algorithms for ASR can be measured in an objective and repeatable way.

1.3.2 *Perceptual evaluation*

Perceptual evaluation, i.e., evaluation using listening tests, requires a group of human assessors to evaluate the processed speech signals. Such evaluation can be conducted to evaluate several attributes, among which speech intelligibility and speech quality are the ones most often considered when evaluating speech enhancement algorithms. Speech intelligibility can be assessed as the number of speech items, i.e., phonemes or words, correctly identified by assessors in relation to the total number of items present in the signal under test. Speech intelligibility is often reported using the speech reception threshold (SRT), which is defined as the level of degradation for which only 50 % of the speech items are correctly identified by an assessor [154,155]. Several frameworks are available to measure the SRT, e.g., based on so-called matrix sentence tests [156]. Provided that the same framework is used, it is possible to measure the SRT for a given assessor multiple times with little variation. Moderate levels of interference may not reduce speech intelligibility while still being detrimental to the end user. Consequently, speech enhancement algorithms may be designed focusing, instead, on the improvement of speech quality.

Contrary to speech intelligibility, speech quality is highly subjective in nature, such that different assessors may have different criteria of what constitutes good or poor speech quality. Consequently, speech quality ratings can greatly differ between assessors and even a single assessor might not consistently assign the same rating to a given stimulus, hence reducing the repeatability of a speech quality test [38]. Moreover, numerous other factors influence speech quality, e.g., noise and reverberation as well as artefacts and distortions introduced by speech enhancement algorithms or hardware limitations. In order to assess the influence of speech enhancement algorithms on speech quality while being conscious of these limitations, results of listening tests are typically averaged over multiple assessors and stimuli. These tests can be broadly classified in two categories, namely relative preference tests and absolute category rating tests.

Speech quality evaluation using relative preference tests relies on paired comparisons of stimuli using either forced-choice comparison (FCC) or comparison category rating (CCR). When using FCC, a pair of stimuli is presented to the assessors who are forced to select the stimulus with the highest perceived speech quality. An FCC-based test provides results in terms of how many percent of time a given algorithm is preferred over an other. However, the magnitude of this difference is not measured. When aiming at quantifying this preference, a CCR-based test should be used, where the preference has to be quantified, typically using a four-point scale [157,158]. A disadvantage of relative preference tests, using either FCC or CCR, is the fact that it can be difficult to compare a large number of algorithms, such that in this case absolute category rating tests are typically used instead.

Speech quality evaluation using absolute category rating tests relies on assessors grading the quality of stimuli on a finite scale. One of the most widely used tests is the mean opinion score (MOS) test. During a MOS test, assessors are instructed to evaluate the quality of each stimulus separately, using a five-point scale labeled from *bad* to *excellent* [36, 157]. As the five levels of quality might not necessarily

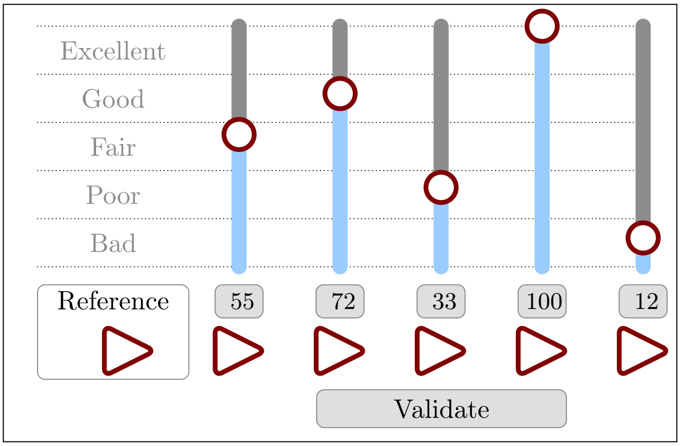

Fig. 1.9: Example of a graphical user interface used to conduct a MUSHRA test

be uniformly spaced, some variants of the MOS test use a continuous scale, where assessors express their ratings as real numbers between zero and ten [159]. Other variants aim at evaluating speech quality indirectly, by requiring assessors to quantify other attributes. For example, the degradation mean opinion score (DMOS) test relies on assessors evaluating the level of degradation of a processed signal, in comparison to its unprocessed counterpart, on a five-point scale labeled from *inaudible* to *very annoying* [160]. DMOS tests are mostly used when the level of degradation is expected to be small. Other measures, such as the diagnostic acceptability measure (DAM), estimate the speech quality by combining the ratings that assessors provide on numerous distinct attributes [161]. DAM is particularly useful if insight into the decision criterion of the assessors is needed but, as it is very time consuming, is avoided if only the overall speech quality is of interest.

In order to evaluate a large amount of algorithms while keeping the test duration manageable, the multiple stimuli test with hidden reference and anchor (MUSHRA) is a popular alternative [37]. An example of a graphical user interface used for such a MUSHRA test is depicted in Fig. 1.9. For each reference, i.e., clean signal under test, assessors are presented with multiple versions of this signal. These variations differ, e.g., in the types of interference that are present in the signals or in the speech enhancement algorithms that have been applied. For each stimulus, assessors use sliders to assign a value between 0 and 100 to a given attribute, e.g., speech quality. The reference signal is presented twice, once labeled and once hidden among the stimuli. The value assigned to this hidden reference can be used to assess the validity of the ratings. One or multiple so-called anchor signals are presented among the stimuli. When measuring speech quality, these anchor signals are typically designed to exemplify poor speech quality and, while a high rating is expected for the hidden reference, low ratings are expected for the anchors. Though originally aimed at evaluating audio codecs, MUSHRA tests are often used to evaluate and compare speech enhancement algorithms [162].

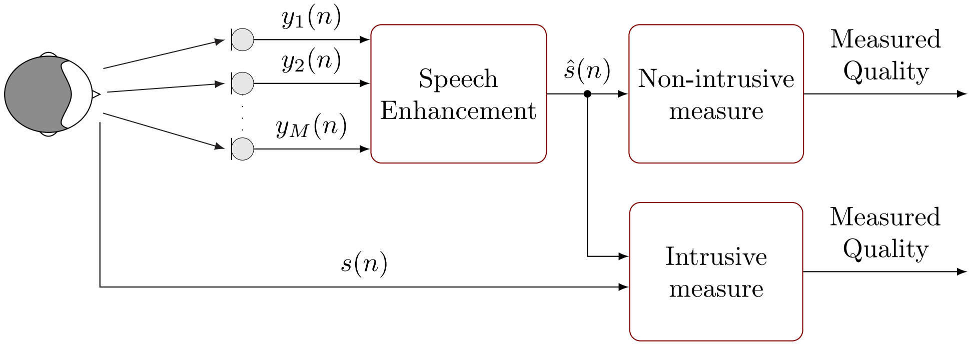

Fig. 1.10: Overview of signal-based speech quality measures. Contrary to intrusive measures, non-intrusive measures can be computed without requiring the clean speech signal.

Regardless of conducted listening test, the significance of the results is typically asserted through statistical analysis, e.g., using analysis of variance (ANOVA) or a signed-rank test. In order for the results to be statistically significant, listening tests may require many assessors, hence being both expensive and time consuming. Consequently, although listening tests remain the gold standard for the evaluation of speech enhancement algorithms, signal-based measures are often the only measures used when developing new algorithms.

1.3.3 *Signal-based measures*

Signal-based measures, often referred to as objective measures, can be categorized as either intrusive or non-intrusive. As depicted in Fig. 1.10, intrusive measures require a reference signal in addition to the test signal, while non-intrusive measures can be computed from the test signal only. Signal-based measures have been proposed to predict speech intelligibility as well as speech quality. Typically, a signal-based measure is considered reliable if it highly correlates with the perceptual ratings [40].

Most existing signal-based measures are intrusive. Among these intrusive measures, the articulation index (AI) [163], the speech transmission index (STI) [164], the speech intelligibility index (SII) [165], the short-time objective intelligibility (STOI)measure [166] and mutual-information-based measures [167], aim at estimating speech intelligibility. Other intrusive measures have been designed to estimate speech quality. The most straightforward measure is the SNR, i.e., the ratio between the power of the signal of interest and the power of the interference. In practice, this ratio is usually computed over short signal segments and the values are limited to avoid the need for a silence detector, e.g., in the computation of the segmental SNR (SSNR) [168]. The SNR can also be computed in the frequency domain, resulting in the so-called frequency-weighted segmental SNR (FWSSNR) [169]. This approach allows to apply larger weighting coefficients to the frequency bands of interests, e.g., based on the articulation index [170]. It has also been proposed to use linear predictive coding (LPC) to quantify the difference between the reference signal and the test signal, e.g., in the log-likelihood ratio (LLR), the Itakura-Saito measure or spectral distortion (SD) [171, 172].

The reliability of measures can typically be improved by using knowledge about the human auditory system. For example, the non-uniformity of the ear's frequency resolution and the nonlinear relationship between the intensity of a sound and its perceived loudness can be taken into account by using the Bark loudness scale [173]. The Bark spectral distortion (BSD) and the modified Bark spectral distortion (MBSD) quantify the difference between the Bark representation of the test and reference signals [174,175]. The perceptual evaluation of speech quality (PESQ) measure [176] is a widely used intrusive measure that as well relies on the Bark scale. PESQ differs from BSD and MBSD, most noticeably by the way in which the difference between the Bark representation of the test and reference signals is computed and by filtering the input signals to mimic the properties of a telephone handset. This filtering is due to the fact that, though widely used in the literature to evaluate speech enhancement algorithms, PESQ was originally designed only to assess the limited amount of distortion caused by telephone transmission and codecs. The perceptual objective listening quality assessment (POLQA) measure [177], that uses a scale analogous to the Bark scale and is often presented as the successor to PESQ, was as well developed for telephone transmission rather than for the evaluation of speech enhancement algorithms. Other intrusive measures have been proposed which make use of more advanced auditory models, such as the perceptual similarity measure (PSM) and, exploiting the finer temporal structure of the test signal, the $\mathrm{PSM}_{\mathrm{t}}$ based on the perception model for quality (PEMO-Q) [178]. In addition, hybrid measures have been proposed, which aim at combining the advantages of different intrusive measures. Hybrid measures are typically based on linear or polynomial regression, such as the multivariate adaptive regression splines (MARS) technique [179]. Although such approaches can improve the correlation with speech quality measured through listening tests, the need for a reference signal still limits the applicability of these measures. In most realistic applications, non-intrusive measures are needed.

Non-intrusive measures, i.e., measures that do not require a reference signal, have been proposed for both speech intelligibility and speech quality estimation. Measures for speech intelligibility include a non-intrusive extensions of the STOI [180, 181], and measures using a trained speech recognizer as proposed in [182, 183]. Similarly as for their intrusive counterparts, non-intrusive measures of the speech quality have not been explicitly developed for the evaluation of speech enhancement algorithms but rather for the evaluation of narrow-band speech codecs. This is for example the case for P.563 [184], which is often described as a non-intrusive alternative to PESQ. In addition, measures such as the speech-to-reverberation modulation energy ratio (SRMR) [185] and its extension, the normalized SRMR ($\mathrm{SRMR}_{\mathrm{norm}}$) [186], have been developed for both speech intelligibility and speech quality and apply a straightforward predicting function to a set of time-averaged modulation energies. Alhough $\mathrm{SRMR}_{\mathrm{norm}}$ has been shown to be promising, e.g., when evaluating cochlear implant processing [187], it can be unreliable when evaluating different types of speech enhancement algorithms [188–190]. Rather than removing the need for a reference signal, it has been proposed in [191] to use twin HMMs to generate an estimate of the clean speech signal before using this estimate and the test signal as input to an intrusive measure. Estimating the clean speech signal is however not straightforward and this method does not outperform the used intrusive measures. Recently promising approaches to develop non-intrusive measures based on machine learning have been proposed. Some measures use machine learning to estimate high-level parameters, which are then combined using more conventional techniques. For example, ANIQUE+ [192] uses deep learning to model distortion and the measure proposed in [193] uses models of clean and degraded speech trained using, e.g., Gaussian mixture models (GMMs). The measure proposed in [194] computes the predicted speech quality as the output of a classification and regression tree (CART) trained using intrusive measures. These measures can be further improved by combining auditory-inspired features with machine learning techniques able to better model the factors influencing the perceived quality of processed speech.

1.4 Outline of the thesis and main contributions

This thesis deals with the **development of speech enhancement algorithms and performance measures to evaluate the quality of speech signals processed by such algorithms**. An important aspect is that the developed algorithms and speech quality measures should be applicable in realistic scenarios. This means that the speech enhancement algorithms should be applicable in real-time and for acoustic scenarios with both noise and reverberation. In addition, the speech quality measures should be non-intrusive, i.e., they can be applied using only the processed signal without the need for a clean reference signal. A structured overview of the thesis is depicted in Fig. 1.11.

The main contributions in this thesis are threefold. First, we have developed **a novel single-channel speech enhancement algorithm for joint noise and reverberation reduction**. The proposed algorithm uses a spectral gain that combines a statistical room acoustics model, minimum statistics and temporal cepstrum smoothing. We have evaluated this algorithm in terms of ASR performance and in terms of speech quality improvement using both signal-based performance measures and a subjective listening test. The results show that the proposed algorithm is able to improve speech quality while reducing reverberation. Moreover, this algorithm yielded the best performance in terms of subjective speech quality among all single-channel algorithms submitted to the REVERB Challenge [31]. Second, we have proposed **two novel non-intrusive speech quality measures that combine perceptually motivated features and predicting functions based on machine learning (model tree, recurrent neural network (RNN))**. The model tree uses time-averaged modulation energies as input, whereas the RNN uses time-varying modulation energies as input in order to take the time-dependency of the test signal into account. The predicting functions used in both proposed speech quality measures have been trained and evaluated using **a corpus of perceptually evaluated speech signals** that we collected for this purpose. This corpus includes a wide range of acoustic scenarios and categories of speech enhancement

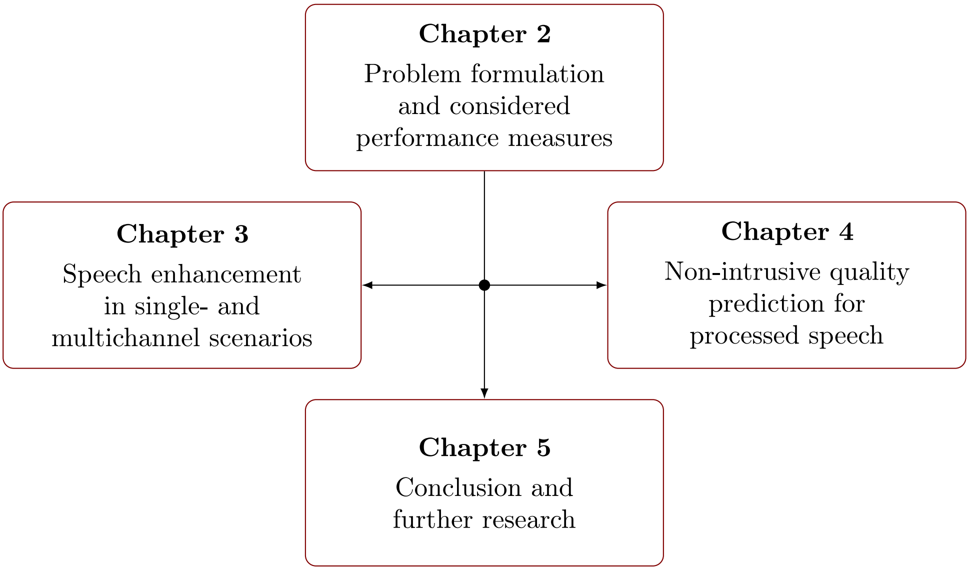
Chapter 2
Problem formulation and considered performance measures
Chapter 3
Speech enhancement in single- and multichannel scenarios
Chapter 4
Non-intrusive quality prediction for processed speech
Chapter 5
Conclusion and further research

Fig. 1.11: Overview of the thesis.

algorithms. Results show that the speech quality measure using an RNN as predicting function outperforms state-of-the-art non-intrusive measures and even yields a similar performance as intrusive measures, when the RNN is trained and tested for a single category of speech enhancement algorithms.

Chapter 2 describes the typical issues of speech enhancement in the presence of noise and reverberation and introduces the notation used throughout the thesis. In addition, this chapter presents the database as well as the speech quality measures and the ASR system used to evaluate the speech enhancement algorithms proposed in Chapter 3. The chapter concludes by presenting the benchmark used to evaluate the speech quality measures proposed in Chapter 4.

In **Chapter 3** we propose a speech enhancement algorithm for joint noise and reverberation reduction. The main contribution is the single-channel spectral enhancement algorithm that uses a spectral gain to enhance the input signal. This spectral gain is computed by combining a statistical room acoustics model, minimum statistics and temporal cepstrum smoothing. This algorithm can be easily combined with existing multichannel algorithms. In this thesis, it is combined with an MVDR beamformer with an online estimated noise coherence matrix and steered towards the DOA of the target speaker, which is estimated using the MUSIC algorithm. We examine the performance of an ASR system, trained on either clean or multi-condition training data, to which we input either unprocessed signals or signals processed using the proposed algorithm. The improvement in speech recognition accuracy is largest when the proposed single-channel speech enhancement algorithm is combined with the MVDR beamformer. When using clean training data, this combination yields an absolute improvement in recognition accuracy of up to 42.24% among the considered acoustic scenarios. When using multi-condition training data, the same combination yields an absolute improvement in recognition accuracy of up to 13.10%. In addition, we examine the performance of the proposed algorithms using signal-based speech quality measures, intrusive measures (PESQ, PEMO-Q) as well as non-intrusive measures (ANIQUE+, P.563, $\mathrm{SRMR}_{\mathrm{norm}}$). With the exception of P.563, all measures show that the proposed algorithms improve speech quality, with the largest improvement coming from the combination of the MVDR beamformer with the single-channel speech enhancement. We also present the results of a MUSHRA test, conducted to assess the speech quality as well as the amount of reverberation in the signals processed with the proposed single-channel spectral enhancement algorithm, the MVDR beamformer, or the combination of the MVDR beamformer with the single-channel spectral enhancement algorithm. For all acoustic scenarios considered in this MUSHRA test, the results show that the proposed algorithms are beneficial in terms of both speech quality and reverberation suppression. This chapter is partly based on the work published in [195, 196].

In **Chapter 4** we propose two non-intrusive speech quality measures that combine perceptually motivated features and predicting functions based on machine learning techniques. The first measure uses time-averaged modulation energies as input to a model tree. The second measure uses time-varying modulation energies as input to an RNN and can hence take the time-dependency of the test signal into account. Both measures are trained and evaluated using a dataset of perceptually evaluated signals that comprises a wide range of acoustic scenarios and categories of algorithms. These algorithms include single-channel spectral enhancement, MVDR beamforming and a combination of MCLP and MVDR beamforming. Using cross-validation, the performance of both measures is benchmarked against intrusive measures (PESQ, PEMO-Q, POLQA) as well as non-intrusive measures (ANIQUE+, P.563, $\mathrm{SRMR}_{\mathrm{norm}}$). The most promising of the proposed speech quality measures is the one using an RNN as predicting function. When trained and tested for a single category of algorithms, this measure outperforms the considered non-intrusive measures and yields a similar performance as the considered intrusive measures. When trained and tested for several categories of algorithms, it even outperforms the intrusive benchmark measures. This chapter is partly based on the work published in [197, 198].

Chapter 5 concludes the thesis by summarizing its main contributions and discussing possible extensions.

2

PROBLEM FORMULATION AND CONSIDERED PERFORMANCE MEASURES

This chapter describes the signal model used throughout the thesis and the corresponding notations in both time and frequency domain. It introduces the objectives guiding the development of speech enhancement algorithms and speech quality measures as well as the benchmarks used to evaluate them. The time and time-frequency domain models for speech signals recorded in a reverberant and noisy environment are described in Section 2.1 before describing the benchmark used to evaluate speech enhancement algorithms in Section 2.2 and the benchmark used to evaluate speech quality measures in Section 2.3.

2.1 Problem formulation

As described in Section 1.1, this thesis considers scenarios in which a single speech source is recorded in a noisy and reverberant environment. In such scenario, the time-domain signal $y_m(n)$ recorded in the m-th microphone of $M \geq 1$ available microphones can, as depicted in Fig. 2.1, be modeled as

$$y_m(n) = x_m(n) + v_m(n) = s(n) * h_m(n) + v_m(n), \tag{2.1}$$

where n denotes the sample index, $s(n)$ denotes the anechoic speech signal, $h_m(n)$ denotes the RIR of length L_h between the source and the m-th microphone and $v_m(n)$ denotes the additive noise component. The reverberant speech component $x_m(n)$ can be written as

$$x_m(n) = d_m(n) + r_m(n), \tag{2.2}$$

where

$$d_m(n) = s(n) * h_m^d(n), \tag{2.3}$$

$$r_m(n) = s(n) * h_m^r(n), \tag{2.4}$$

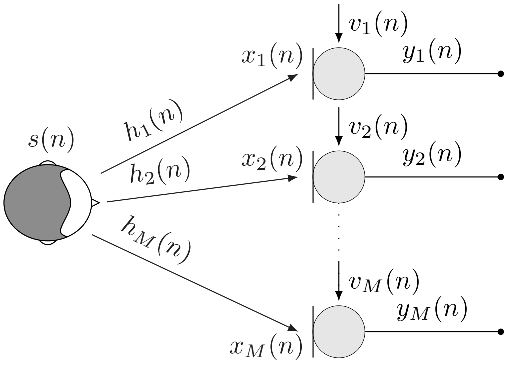

Fig. 2.1: Signal model for a single speech source recorded in a noisy and reverberant environment.

with $h_m^d(n)$ and $h_m^r(n)$ defined as

$$h_m^d(n) = \begin{cases} h_m(n) \text{ for } n \leq L_d, \\ 0 \text{ otherwise}, \end{cases} \tag{2.5}$$

$$h_m^r(n) = \begin{cases} h_m(n) \text{ for } n > L_d, \\ 0 \text{ otherwise}, \end{cases} \tag{2.6}$$

where L_d is set so that $h_m^d(n)$ contains the direct path and a few early reflections while $h_m^r(n)$ contains the late reflections, i.e., the reverberant tail. The output signal $\hat{s}(n)$ of a speech enhancement algorithm is computed from the recorded microphone signals $y_m(n)$, $m \in [1 \cdots M]$, as an estimate of either $s(n)$ or $d_{\text{ref}}(n)$, where ref denotes the index of a reference microphone.

The algorithms considered in this thesis process the recorded signal in the time-frequency domain by using the STFT. The STFT of the input signal is obtained by computing the discrete Fourier transform (DFT) of overlapping frames of the time-domain signal weighted by an analysis window $w_{\text{STFT}}(n)$, i.e.,

$$y_m(k, \ell) = \sum_{n=0}^{N-1} w_{\text{STFT}}(n) y_m(\ell R + n) \mathrm{e}^{\frac{-j2\pi kn}{N}}, \tag{2.7}$$

where k and ℓ denote the frequency bin and frame index, respectively, N denotes the frame length and R denotes the frame shift. The time-domain model from (2.1) and (2.2) can be written in the STFT domain as

$$y_m(k,\ell) = x_m(k,\ell) + v_m(k,\ell), \tag{2.8}$$

$$x_m(k,\ell) = d_m(k,\ell) + r_m(k,\ell), \tag{2.9}$$

where $x_m(k,\ell)$, $v_m(k,\ell)$, $d_m(k,\ell)$ and $r_m(k,\ell)$ denote the STFTs of $x_m(n)$, $v_m(n)$, $d_m(n)$ and $r_m(n)$, respectively. Speech enhancement in the STFT domain computes an estimate $\hat{s}(k,\ell)$, of either $s(k,\ell)$, i.e., the STFT of $s(n)$, or $d_m(k,\ell)$. The time-domain output of the speech enhancement algorithm can finally be computed using the ISTFT as

$$\hat{s}(n) = \sum_{\ell} \sum_{k=0}^{N-1} w_{\text{ISTFT}}(n - \ell R)\hat{s}(k,\ell)\mathrm{e}^{\frac{j2\pi k(n-\ell R)}{N}}, \tag{2.10}$$

where $w_{\text{ISTFT}}(n)$ denotes a synthesis window that, along with $w_{\text{STFT}}(n)$, usually aims at satisfying the perfect overlap-add constraint [199]. A benchmark is needed to assess the ability of such speech enhancement algorithms to improve the performance of ASR systems or the perceived speech quality. The next section describes the benchmark used to evaluate the speech enhancement algorithms considered in this thesis.

2.2 Speech enhancement benchmark

Though numerous speech enhancement algorithms have been proposed, their evaluation and their comparison with prior work is often limited due to the lack of a publicly available benchmark. This is particularly true for algorithms designed to improve speech quality in noisy and reverberant environments as most existing benchmarks aimed at evaluating only robustness against noise and were often designed to evaluate only the benefits of noise reduction algorithms or the performance of ASR systems [34, 200, 201]. The speech enhancement algorithm, proposed in this thesis and presented in Chapter 3, is evaluated using the corpus from the REVERB challenge [31, 162]. This corpus is designed to evaluate speech enhancement algorithms in presence of reverberation and moderate level of noise when using a single-channel, two channels or the eight channels of a circular microphone array with a 20 cm diameter. Only the single-channel and eight channels scenarios are considered in this thesis. This corpus comprises a dataset of audio material to be processed by the speech enhancement algorithms to be evaluated as well as evaluation tools. The dataset, divided into a development set and an evaluation set, comprises both simulated and real data, in which all signals have a sampling frequency f_s of 16 kHz. An overview of this dataset is depicted in Fig. 2.2.

The simulated data is generated by convolving clean speech from the WSJCAM0 corpus [202] with measured RIRs and adding ambient noise to this reverberant signal. The level of the noise is scaled to obtain an SNR of 20 dB measured considering

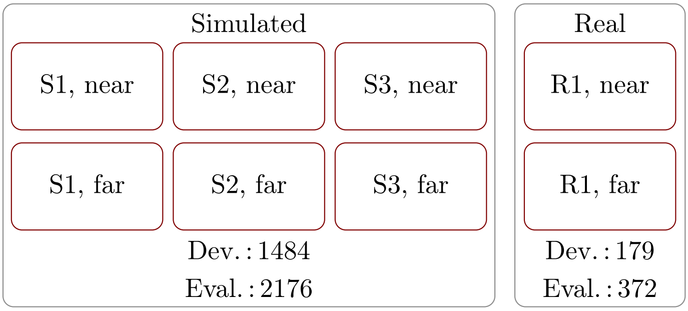

Fig. 2.2: Overview of the REVERB challenge dataset and number of utterances of either simulated or real signals present in the development (Dev.) and evaluation (Eval.) sets.

the early reverberant speech as target signal and ignoring speech pauses. For each recorded RIR, the used noise was recorded using the same microphone position. This ambient noise was mainly due to the air conditioning system and, consequently, is quite stationary and has most of its energy present in the lower frequency. Though more challenging noise conditions could be considered, this dataset focuses on the evaluation of speech enhancement algorithms in presence of both noise and reverberation. The simulated data represents various reverberation conditions and is generated using 6 RIRs, recorded in 3 different rooms (small, medium and large) and considering two distances per room. In the remainder of this thesis, these rooms will be denoted as "S1", "S2" and "S3" while the distances will be referred to as "near" and "far". For each combination of room and distance, e.g., "S2, near", the distance between speech source and microphone array as well as the T_{60} and D_{50} measured from the RIR are summarized in Table 2.1. A detailed description of the simulation and RIRs recording is available in [203].

The real data is composed of recordings of utterances from the MC-WSJ-AV corpus [204] spoken by a human speaker in a noisy and reverberant room. The noise type is similar to the one used to generate the simulated data. In absence of the clean signal, neither the SNR, nor the T_{60} nor the D_{50} can be exactly measured. However, despite a slightly lower SNR and a larger distance between the speech source and the array, the noise and reverberation conditions in the real dataset are similar to the one in the room S3 of the simulated data. Recording distance and estimated T_{60} for the real data are also summarized in Table 2.1.

2.2.1 *Considered ASR system*

The sentences spoken in both simulated and real data are the same and, consequently, the same ASR system can be used to evaluate the impact that speech enhancement algorithms would have on the performance of such systems when applied to the previously described dataset. The ASR system used in this thesis is the one that was provided as baseline for the ASR task of the REVERB Challenge. Details of its design and comparison of its performance with more advanced systems

Table 2.1: Distance (Dist.) between speech source and center of the microphone array as well as T_{60} and D_{50} for all combinations of room and distance included in the considered dataset. For the simulated data, T_{60} and D_{50} were measured from the recorded RIRs. For the real data, the T_{60} was measured in the recording room but D_{50} is not available.

	Simulated						Real	
	S1, near	S1, far	S2, near	S2, far	S3, near	S3, far	R1, near	R1, far
Dist. [cm]	50	200	50	200	50	200	100	250
T_{60} [ms]	300	300	600	600	700	700	700	700
D_{50}	0.99	0.98	0.95	0.79	0.97	0.81		

can be found in [31]. It relies on a bigram scheme language model and on an acoustic model based on GMMs and HMMs.

Within this thesis, ASR performance is measured using two different acoustic models. First, one trained on clean speech from the WSJCAM0 dataset and one, so-called, multi-condition model that has been trained on data generated by convolving the clean speech with 24 measured RIRs with T_{60} ranging from 0.2 s to 0.8 s and adding ambient noise at an SNR of 20 dB. All models were trained using the hidden Markov model toolkit (HTK) [205]. It can be noted that tools to train and evaluate a DNN-based recogniser on the same data is available as part of the Kaldi toolkit [206]. This DNN-based recogniser is not used in this thesis as ASR is here solely used to evaluate and compare speech enhancement algorithms rather than to obtain accurate text transcriptions. The performance of the considered ASR systems will be reported using the recognition accuracy,

$$\text{Accuracy}\,[\%] = 100 - \text{WER}, \tag{2.11}$$

where the word error rate (WER) is defined as

$$\text{WER}\,[\%] = 100 \cdot \frac{\text{substitutions} + \text{deletions} + \text{insertions}}{\text{total}}, \tag{2.12}$$

i.e., the number of words which have been substituted by an other, deleted or inserted, divided by the total number of words to be recognized over the complete considered corpus.

2.2.2 *Considered quality measures*

The speech enhancement algorithms presented in this thesis are evaluated in terms of the resulting speech quality using listening tests based on the MUSHRA test [37] and, in Chapter 3, a similar test is used to measure the perceived level of reverberation. As the exact specifications of each test, e.g., choice of the anchors or permutations of the presented conditions, may vary between tests, they will be presented in each respective chapter. However, the *true* speech quality for a combination of algorithm and acoustic condition is, through the entirety of this thesis, considered to be the mean of the scores assigned by all assessors participating in a given test. More precisely, each processed signal $\hat{s}(n)$ is assigned a quality value $0 \leq p_{\hat{s}} \leq 1$, which is computed by averaging all MUSHRA scores assigned to it and by normalizing this average between 0 and 1. In addition to listening tests, the quality of the processed signals will as well be evaluated using intrusive and non-intrusive signal-based measures.

Three intrusive signal-based measures of the speech quality are considered in this thesis, namely PESQ [176], POLQA [177] and PEMO-Q [178]. The computation of the PESQ score relies on time-aligning the clean reference signal and the test signal before filtering them to replicate the effect of telephone handsets. This filtering is applied because, though very often used in the speech enhancement literature, PESQ was developed for the evaluation of speech codecs. The Bark representation of each signal is extracted and the difference between the Bark representation of the reference signal and test signal is computed. A positive difference indicates that a disturbance, e.g., noise, is present while a negative difference indicates that part of the speech to be preserved has been suppressed in the test signal. The final PESQ score is computed as a linear combination of these negative and positive Bark scale differences. The coefficients used for this combination were optimized using perceptually evaluated telephone data. It might be noted that, though the original version of the PESQ score was limited to signals sampled at 8 kHz, the version used in this thesis is the later extension for signals having a sampling frequency f_s of 16 kHz. POLQA has been developed to supersede PESQ and to be applicable to signals sampled at higher frequencies and containing a wider range of distortions. However, due to licence limitations of POLQA, i.e., its source code is not available and its licence is costly, PESQ remains more widely used for speech enhancement assessment. In this thesis, POLQA scores are not available for the speech enhancement algorithms presented in Chapter 3 but are part of the benchmark used to develop a non-intrusive quality measures in Chapter 4. Finally, PEMO-Q aims at comparing auditory-inspired representations of the test and of the reference signal. This auditory representation is obtained by using a gammatone filter bank and filtering each resulting band using a low-pass filter and a combination of feedback loops. The output of this processing chain is finally analysed by a linear modulation filter bank whose output is used to compute the difference between the reference and the test signal.

Three non-intrusive signal-based measures of the speech quality are considered in this thesis, namely ANIQUE+ [192], P.563 [184] and $\text{SRMR}_{\text{norm}}$ [186]. These non-intrusive measures are used to evaluate speech enhancement algorithms in Chapter 3 as well as to benchmark quality measures in Chapter 4. ANIQUE+ computes the estimated speech quality by combining scores assigned to three types of effects that its authors refer to as *frame distortion*, *mute impact* and *non-speech impact*. Though the quantification of each of these criteria relies on similar preprocessing and filterbank, each of them requires a separate model that needs to be carefully tuned. So far, ANIQUE+ is often included when, as in this thesis, one aims at benchmarking its own measure, but is seldom applied to the evaluation of speech enhancement algorithms. P.563 is sometimes presented as the non-intrusive counterpart of PESQ

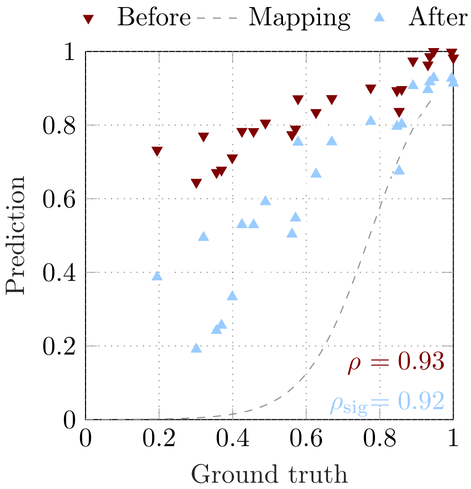

Fig. 2.3: Example of predicted value of the speech quality, before and after applying the sigmoidal mapping used to compute ρ_{sig}. The values of ρ and ρ_{sig} in this example show that this values may be similar and that the mapping does not necessarily increase correlation.

and has long been considered the state-of-the-art of non-intrusive speech quality estimation [184, 193]. The P.563 score is computed by, first, detecting one of six possible types of distortions such as *unnatural male or female speech* or *high level of background noise*. The intermediate quality ratings assigned to these distortions are linearly combined to compute the final quality estimate. It may be noted that P.563 has been developed to assess the quality of speech transmitted over telephone lines and that, therefore, it only considers distortions present in this use case. In addition, it is only applicable to narrow-band speech signals and, as in this thesis, signals having a higher sampling frequency have to be downsampled to 8 kHz before applying P.563. The $\text{SRMR}_{\text{norm}}$ is an extension of the SRMR initially proposed in [185] as a non-intrusive quality measure for reververberant and dereverberated speech. Both SRMR and $\text{SRMR}_{\text{norm}}$ are computed similarly. First, the test signal is filtered through a gammatone filterbank whose characteristics are empirically defined for signals recorded at either 8 kHz or 16 kHz. Then, the temporal envelope of each band is computed using the Hilbert transform and modulation energies are computed. SRMR and $\text{SRMR}_{\text{norm}}$ differ in the settings applied in those steps and more details are provided in Chapter 4. As the settings of $\text{SRMR}_{\text{norm}}$ have been shown to improve the reliability of the measure, they are the ones that are used throughout this thesis. The final quality estimation of either SRMR or $\text{SRMR}_{\text{norm}}$ is finally computed as the ratio between the energies in the higher and lower modulation frequencies after averaging them over time.

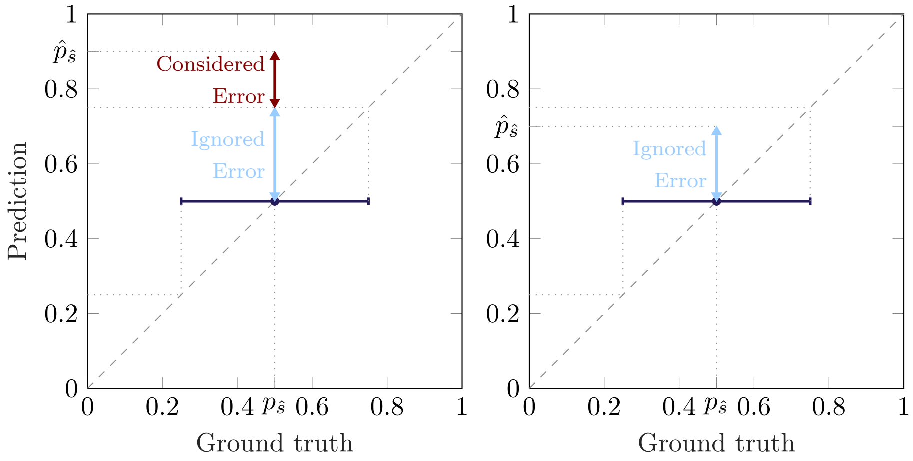

Fig. 2.4: Error thresholding to compute ϵ-RMSE, the horizontal error bar represents the 95% confidence interval of the subjective test used to obtain $p_{\hat{s}}$. Left, the total error is larger than the confidence interval and their difference is considered as error to compute the ϵ-RMSE. Right, the total error is smaller than the confidence interval and all error is ignored.

2.3 Quality measures benchmark

Signal-based quality measures, either intrusive or non-intrusive, are only useful if they yield a reliable estimate $\hat{p}_{\hat{s}}$ of what is considered the true quality $p_{\hat{s}}$ of the processed signal $\hat{s}(n)$ under test. In this thesis, this reliability is assessed by examining the correlation between the considered quality measures and the true speech quality as well as the spread of these measures. Three different values of correlation are considered. The first one is the Pearson correlation defined as

$$\rho = \frac{\sum(p_{\hat{s}} - \mu_{p_{\hat{s}}})(\hat{p}_{\hat{s}} - \mu_{\hat{p}_{\hat{s}}})}{\sqrt{\sum(p_{\hat{s}} - \mu_{p_{\hat{s}}})^2}\sqrt{\sum(\hat{p}_{\hat{s}} - \mu_{\hat{p}_{\hat{s}}})^2}}, \tag{2.13}$$

where all sums as well as the means $\mu_{p_{\hat{s}}}$ and $\mu_{\hat{p}_{\hat{s}}}$ are computed over all values of $p_{\hat{s}}$ and $\hat{p}_{\hat{s}}$ available in the considered dataset. Second, the Spearman correlation ρ_{spear} is computed similarly as in (2.13) by substituting the values of $p_{\hat{s}}$ and $\hat{p}_{\hat{s}}$ by their rank among the considered dataset. Finally, the sigmoidal correlation ρ_{sig} is computed similarly as in (2.13) after applying a sigmoidal mapping to the values of $\hat{p}_{\hat{s}}$ available in the considered dataset. The parameters of this mapping have to be learned from a training set for which the values of both $p_{\hat{s}}$ and $\hat{p}_{\hat{s}}$ are known for a large number of signals $\hat{s}(n)$. When computing ρ_{sig} to evaluate quality measures based on machine learning techniques in Chapter 4, the same training set is used to train the measures and determine the parameters of the sigmoidal mapping. An example of such mapping is depicted in Fig. 2.3. The value of ρ_{sig} can be lower than the value of ρ. The main appeal of using such mapping is that it allows to normalize the range of the predicted scores in a less arbitrary way than, e.g., when simply dividing by the largest measured value.

Along with the correlation, an important aspect to consider when evaluating a speech quality measure is the spread of its prediction. This spread represents the ability of the measure to assign similar scores to signals having the same *true* quality, i.e., lower spreads are better. Spreads are often measured using the root mean square error (RMSE) that is computed from the difference between ground truth and prediction for a set of observations. However, in the case of speech quality measures, the ground truth itself is based on averaging the scores assigned by several participants during a listening test. As subjective scores themselves can vary, one might want to take this uncertainty into account when evaluating speech quality measures. For this purpose, the so-called epsilon insensitive RMSE (ϵ-RMSE) is used in this thesis. This measure is similar to the traditional RMSE but measures the error from the sigmoidally mapped scores only and set it to zero if this error is lower than the 95% confidence interval of the subjective scores. Fig. 2.4 summarizes the way in which this error is computed but further details on the computation of ρ_{sig} and ϵ-RMSE can be found in [40].

2.4 Summary

In this chapter, we presented the models and main notations used through this thesis to described noisy and reverberant signal in both time and time-frequency domain. We introduced the corpus used to benchmark the speech enhancement algorithms, that includes both real and simulated data. We briefly described the tools used to built ASR systems trained on clean and multi-condition data to evaluate the benefits that the speech enhancement algorithms considered in this thesis could have on ASR performance. We presented the signal-based measures of speech quality, intrusive and non-intrusive, that are used in this thesis to evaluate speech enhancement algorithms in Chapter 3 as well as to benchmark the proposed speech quality measures in Chapter 4. Finally, we introduced the figures of merits used to benchmark these speech quality measures.

3

SPEECH ENHANCEMENT IN SINGLE- AND MULTICHANNEL SCENARIOS

In many speech communication applications, such as voice-controlled systems or hearing aids, distant microphones are used to record a target speaker. The microphone signals are often corrupted by both reverberation and noise, resulting in a degraded speech quality and speech intelligibility, as well as in a reduced performance of ASR systems. Several algorithms have been proposed in the literature to deal with these issues and are often based on combinations of beamforming and single-channel spectral enhancement, as introduced in Section 1.2.

This chapter presents the algorithm first proposed in [195] and further evaluated in [196], that consists of the commonly used combination of an MVDR beamformer with a single-channel spectral enhancement algorithm. In such a combined algorithm, the spectral enhancement algorithm typically consists in applying a real-valued spectral gain to the STFT of the beamformer output. The computation of this spectral gain relies on estimates of the PSDs of the interferences to be suppressed, i.e., noise and reverberation. In this case, the reverberation to be suppressed consists only of the late reflections as early reflections are often considered to be beneficial, both in terms of speech quality [9] and ASR performance [207].

Different methods have been proposed for estimating the late reverberant and noise PSDs, e.g., relying on assumptions about the sound field or on a VAD. The PSDs of the noise and reverberation can be estimated using the output signal(s) of a blocking matrix, suppressing the signal to be preserved, in the well-known generalized sidelobe canceller (GSC) structure. The blocking matrix can be designed, e.g., as a

---

This chapter is partly based on:

[195] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech-enhancement scheme," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014

[196] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 61, pp. 1–12, Jul. 2015

delay-and-subtract beamformer cancelling the direct speech component [208,209] or based on a blind source separation (BSS) scheme aiming to cancel both the direct speech component and the early reflections [210,211]. Alternatively, the PSD at a reference position can be obtained using a maximum likelihood estimator and a model of the sound field [82]. The PSD to be used in the computation of the spectral postfilter is then obtained by correcting the estimated PSD at the reference position. This correction can be done using an adaptive filter [209], back-projection [210, 211] or the relative transfer functions between the target speaker and the microphones [82].

Other algorithms estimate the PSD of the interference from the output of the beamformer and thus can in principle also be used if only one microphone is available. In such algorithms [195,196,212], the estimation of the noise PSD is often derived from statistical models of the speech and noise [90,94]. The estimation of the reverberant PSD can, e.g., be derived from a statistical model of the RIR and the acoustical properties of the room, such as the reverberation time (T_{60}) or the DRR [81, 87].

In the algorithm presented in this chapter, the microphone signals are first processed using an MVDR beamformer [213], which aims at suppressing sound sources not arriving from the DOA of the target speaker, while maintaining a unit gain towards this DOA. The noise coherence matrix used to compute the coefficients of the MVDR beamformer is estimated online using a VAD [89], and the DOA of the target speaker is estimated using the MUSIC algorithm [117, 118]. The beamformer output is processed using a single-channel spectral enhancement algorithm, that aims at jointly suppressing the residual noise and reverberation. The main novel contribution in [195, 196], on which this chapter is partially based, is the combination of several estimators used in the single-channel spectral enhancement algorithm. This spectral enhancement algorithm relies on estimates of the PSDs of the noise and the late reverberation, similarly as in [214]. The proposed algorithm computes a real-valued spectral gain, combining the clean speech amplitude estimator presented in [215], the noise PSD estimator based on MS [90], and an estimator of the (late) reverberant PSD based on statistical room acoustics [24, 87]. In order to reduce the musical noise which is often a byproduct of spectral enhancement algorithms, adaptive smoothing in the cepstral domain is used to estimate the speech PSD [216, 217].

The proposed algorithm is evaluated on the REVERB challenge corpus, described in Chapter 2. The single-channel scenario is particularly challenging as illustrated by the results of the REVERB challenge workshop [31,162], in which most contributions succeeded to reduce reverberation but only a few improved the speech quality [195, 212]. The evaluation is conducted for different configurations of the proposed system in terms of instrumental speech quality measures, improvement of ASR performance and a subjective evaluation of speech quality and dereverberation. The evaluation results show that the proposed system is able to reduce noise and reverberation while improving the speech quality in both single- and multichannel scenarios.

The remainder of this chapter is organized as follows. In Section 3.1, an overview of the considered system is given as well as details of the proposed MVDR beamformer. The single-channel spectral enhancement algorithm is presented in Section 3.2. The experiments are described in Section 3.3 and the results are presented in Section 3.4.

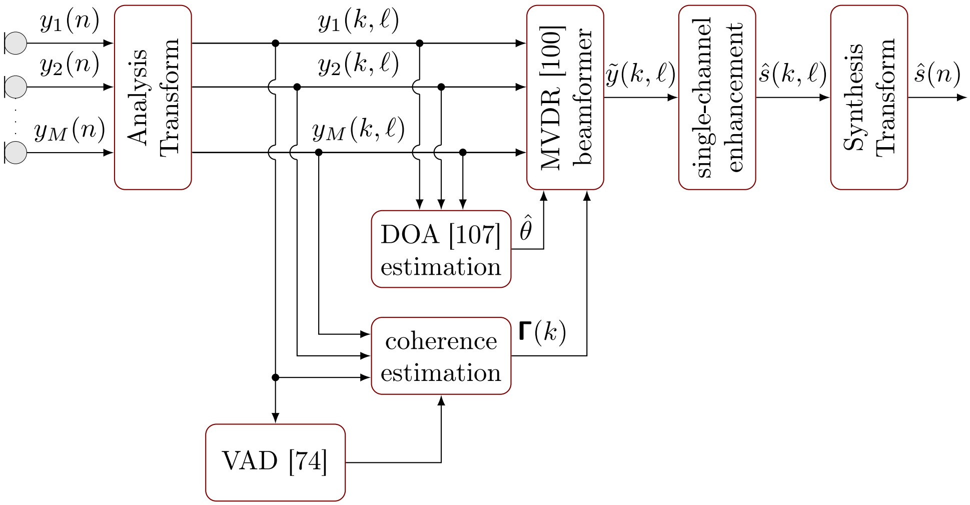

Fig. 3.1: Overview of the proposed algorithm. The STFT of the multichannel input is processed using an MVDR beamformer whose parameters are estimated online. Residual noise and reverberation are suppressed by a single-channel spectral enhancement algorithm (see Fig. 3.2).

3.1 System overview and beamformer

The proposed algorithm, depicted in Fig. 3.1, aims at obtaining an estimate $\hat{s}(n)$, where $\hat{\cdot}$ denotes estimated quantities, of the clean speech signal $s(n)$ from the reverberant and noisy microphone signals, $y_m(n)$. This algorithm consists of two stages. First, an MVDR beamformer is applied to the STFTs $y_m(k,\ell)$ of the M input signals. This beamformer aims at reducing noise and reverberation by suppressing the sound sources not arriving from the target DOA, while providing a unity gain in the direction of the target speaker. The noise coherence matrix and the DOA used to compute the MVDR beamformer coefficients are estimated online, allowing it to be applied in realistic settings. The noise coherence matrix is estimated using a VAD [89], whereas the DOA estimation is based on the MUSIC algorithm [117,118]. In order to suppress the residual noise and reverberation at the beamformer output $\tilde{y}(k,\ell)$, the beamformer output is processed by a single-channel spectral enhancement algorithm. The remainder of this section describes the used MVDR beamformer and the single-channel spectral enhancement algorithm is described in Section 3.2.

3.1.1 *MVDR beamforming*

From the signal model presented in (2.8), the STFT of the multichannel input signal can be written in vector form as

$$\mathbf{y}(k,\ell) = \mathbf{x}(k,\ell) + \mathbf{v}(k,\ell), \tag{3.1}$$

where $\mathbf{y}(k,\ell)$ denotes the M-dimensional stacked vector of the received microphone signals,

$$\mathbf{y}(k,\ell) = [y_1(k,\ell),\ y_2(k,\ell),\ \ldots\ , y_M(k,\ell)]^{\mathrm{T}}, \tag{3.2}$$

and where $\mathbf{x}(k,\ell)$ and $\mathbf{v}(k,\ell)$ denote the similarly defined vectors of the reverberant speech component and noise component, i.e.,

$$\mathbf{x}(k,\ell) = [x_1(k,\ell),\ x_2(k,\ell),\ \ldots\ , x_M(k,\ell)]^{\mathrm{T}}, \tag{3.3}$$

$$\mathbf{v}(k,\ell) = [v_1(k,\ell),\ v_2(k,\ell),\ \ldots\ , v_M(k,\ell)]^{\mathrm{T}}. \tag{3.4}$$

The STFT $\tilde{y}(k,\ell)$ of the beamformer output can be computed as

$$\begin{aligned}\tilde{y}(k,\ell) &= \mathbf{w}_{\hat{\theta}}^{\mathrm{H}}(k)\mathbf{y}(k,\ell) && (3.5)\\ &= \mathbf{w}_{\hat{\theta}}^{\mathrm{H}}(k)\mathbf{x}(k,\ell) + \mathbf{w}_{\hat{\theta}}^{\mathrm{H}}(k)\mathbf{v}(k,\ell) && (3.6)\end{aligned}$$

where $\mathbf{w}_{\hat{\theta}}(k)$ denotes the stacked filter coefficient vector of the beamformer steered towards the estimate $\hat{\theta}$ of the DOA, θ, of the target speech.

Aiming at minimizing the noise power while providing a unity gain in the direction of the target speaker, the filter coefficients of the MVDR beamformer are computed as [213]

$$\mathbf{w}_{\hat{\theta}}(k) = \frac{\hat{\mathbf{\Gamma}}^{-1}(k)\mathbf{d}_{\hat{\theta}}(k)}{\mathbf{d}_{\hat{\theta}}^{\mathrm{H}}(k)\hat{\mathbf{\Gamma}}^{-1}(k)\mathbf{d}_{\hat{\theta}}(k)}, \tag{3.7}$$

where $\mathbf{d}_{\hat{\theta}}(k)$ and $\hat{\mathbf{\Gamma}}(k)$ denote the steering vector towards the estimated DOA of the target speaker and the estimated noise coherence matrix, respectively. Using a far-field assumption, the steering vector $\mathbf{d}_{\hat{\theta}}(k)$ can be computed as

$$\mathbf{d}_{\hat{\theta}}(k) = [e^{-j2\pi f_k \tau_1(\hat{\theta})},\ e^{-j2\pi f_k \tau_2(\hat{\theta})},\ \ldots\ , e^{-j2\pi f_k \tau_M(\hat{\theta})}]^{\mathrm{T}}, \tag{3.8}$$

where f_k denotes the center frequency of the k-th frequency bin and $\tau_m(\hat{\theta})$ denotes the time difference of arrival of the source at angle $\hat{\theta}$ between the m-th microphone and a reference position. In this thesis, this reference position is arbitrarily chosen as the center of the microphone array.

The computation of the estimate $\hat{\theta}$ of the DOA and of the estimate $\hat{\mathbf{\Gamma}}(k)$ of the noise coherence matrix are detailed in the next subsections.

3.1.2 *Noise coherence matrix estimation*

The noise coherence matrix is estimated during noise-only periods detected using the VAD described in [89], as the covariance matrix of the noise-only components, i.e.,

$$\hat{\mathbf{\Gamma}}(k) = \frac{1}{\overline{\overline{\mathbb{L}}}_v} \sum_{\ell \in \mathbb{L}_v} \mathbf{v}(k,\ell)\mathbf{v}^{\mathrm{H}}(k,\ell), \tag{3.9}$$

with $\mathbb{L}_v$ denoting the set of detected noise-only frames and $\overline{\overline{\mathbb{L}}}_v$ its cardinality. However, if the detected noise-only period is too short for a reliable estimate (see Section 3.3), the coherence matrix $\bar{\mathbf{\Gamma}}(k)$ of a diffuse noise field is used instead, i.e., the coherence between two microphones i and i', separated by a distance $l_{i,i'}$, is computed as

$$\bar{\mathbf{\Gamma}}_{i,i'}(k) = \frac{\sin\left(2\pi f_k l_{i,i'}/c\right)}{2\pi f_k l_{i,i'}/c}, \tag{3.10}$$

where c denotes the speed of sound, resulting in the well-known superdirective beamformer [213]. In addition, a white noise gain constraint $\mathrm{WNG}_{\max}$ is imposed in order to limit the potential amplification of uncorrelated noise, especially at low frequencies. With such a constraint, the estimate of the noise coherence matrix used to compute the beamformer coefficients is equal to

$$\hat{\mathbf{\Gamma}}(k) = \bar{\mathbf{\Gamma}}(k) + \varrho(k)\mathbf{I}_M, \tag{3.11}$$

where $\mathbf{I}_M$ denotes the $M \times M$-dimensional identity matrix and $\varrho(k)$ denotes a frequency-dependent regularization parameter which is computed iteratively such that $\mathbf{w}_\theta^{\mathrm{H}}(k)\mathbf{w}_\theta(k) \leq \mathrm{WNG}_{\max}$ [218].

3.1.3 *DOA estimation*

As the beamformer aims at suppressing sources not arriving from the target DOA, an error in the DOA estimate may lead to suppression of the desired source by the beamformer. In the proposed system, the subspace-based MUSIC algorithm [117, 118], shown robust in our target application (see Section 3.3), has been used to compute the DOA estimate $\hat{\theta}$.

Assuming that speech and noise are uncorrelated, the steering vector corresponding to the true DOA is orthogonal to the noise subspace $\mathbf{U}(k,\ell)$. This noise subspace is represented by an $M \times (M-Q)$-dimensional matrix , where Q denotes the number of sources, i.e., $Q = 1$ in this case, and is defined as

$$\mathbf{U}(k,\ell) = [\mathbf{u}_{Q+1}(k,\ell),\ \cdots,\mathbf{u}_M(k,\ell,)]. \tag{3.12}$$

The noise subspace $\mathbf{U}(k,\ell)$ is composed of the eigenvectors of the covariance matrix of $\mathbf{y}(k,\ell)$ corresponding to the $(M-Q)$ smallest eigenvalues. The MUSIC algo-

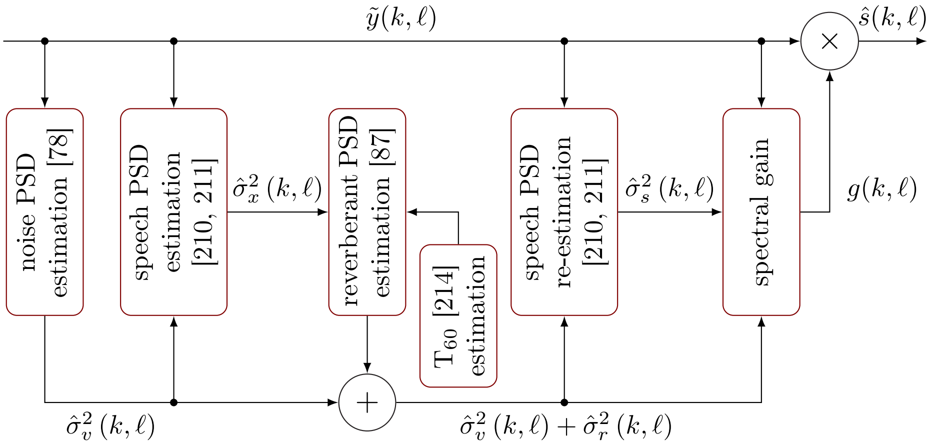
$\tilde{y}(k,\ell)$
$\hat{s}(k,\ell)$
noise PSD estimation [78]
speech PSD estimation [210, 211]
$\hat{\sigma}_x^2(k,\ell)$
reverberant PSD estimation [87]
T$_{60}$ [214] estimation
speech PSD re-estimation [210, 211]
$\hat{\sigma}_s^2(k,\ell)$
spectral gain
$g(k,\ell)$
$\hat{\sigma}_v^2(k,\ell)$
$\hat{\sigma}_v^2(k,\ell)+\hat{\sigma}_r^2(k,\ell)$

Fig. 3.2: Overview of the proposed single-channel speech enhancement.

rithm uses this noise subspace to estimate, for each frame, the DOA as the angle maximizing the sum of the MUSIC pseudo-spectra,

$$\tilde{u}_\theta(k,\ell) = \frac{1}{\mathbf{d}_\theta^{\mathrm{H}}(k)\mathbf{U}(k,\ell)\mathbf{U}^{\mathrm{H}}(k,\ell)\mathbf{d}_\theta(k)}, \tag{3.13}$$

over a given frequency range, i.e.,

$$\hat{\theta}(\ell) = \underset{\theta}{\operatorname{argmax}} \sum_{k=k_{\text{low}}}^{k_{\text{high}}} \tilde{u}_\theta(k,\ell), \tag{3.14}$$

where k_{low} and k_{high} denote the indices of the frequency bin corresponding to the lowest and highest considered frequencies. When assuming, as in (3.8), that the DOA is constant, one can compute $\hat{\theta}$ as the median value of $\hat{\theta}(\ell)$ over all considered frames.

3.2 Single-channel spectral enhancement

Although the beamformer in Section 3.1 is able to reduce the interference, i.e., noise and reverberation, to some extent, spectral enhancement algorithms are able to further reduce reverberation as well as noise. The output signal $\tilde{y}(k,\ell)$ of the MVDR beamformer contains the clean speech signal $s(k,\ell)$ as well as residual reverberation $r(k,\ell)$ and residual noise $v(k,\ell)$, i.e.,

$$\tilde{y}(k,\ell) = x(k,\ell) + v(k,\ell), \tag{3.15}$$

where

$$x(k,\ell) = s(k,\ell) + r(k,\ell), \tag{3.16}$$

denotes the reverberant speech component. Aiming at jointly reducing residual reverberation and noise, the single-channel spectral enhancement algorithm summarized in Fig. 3.2 is proposed, where a real-valued spectral gain $g(k,\ell)$ is applied to the STFT coefficients of the beamformer output, i.e.,

$$\hat{s}(k,\ell) = g(k,\ell)\tilde{y}(k,\ell), \tag{3.17}$$

where $\hat{s}(k,\ell)$ denotes the STFT of the estimated speech signal.

The spectral gain $g(k,\ell)$ is computed using the minimum mean square error (MMSE) estimator for the clean speech spectral magnitude as proposed in [215] (see Subsection 3.2.1). This estimator, similarly to the Wiener filter, requires the PSDs of the clean speech, of the noise and of the reverberation components.

First, the estimate $\hat{\sigma}_v^2(k,\ell)$ of the noise PSD $\sigma_v^2(k,\ell)$ is computed based on a slight modification of the well-known minimum statistics (MS) approach [90] (see Subsection 3.2.2) and used to estimate the reverberant speech PSD. The estimate $\hat{\sigma}_x^2(k,\ell)$ of the reverberant speech PSD $\sigma_x^2(k,\ell)$ is computed using temporal cepstrum smoothing [216,217] (see Subsection 3.2.3). The estimate $\hat{\sigma}_r^2(k,\ell)$ of the (late) reverberant PSD $\sigma_r^2(k,\ell)$ is computed from the reverberant speech PSD estimate using the approach proposed in [87] (see Subsection 3.2.4). This approach requires an estimate of the reverberation time T_{60}, which, in this thesis, is obtained using the estimator described in [219]. As the dereverberation task is treated separately from the denoising task, care has to be taken that no reverberation leaks into the noise PSD estimate and vice versa. Thus, a longer minimum search window is used in the MS approach as compared to [90] (see Section 3.3).

The estimate $\hat{\sigma}_s^2(k,\ell)$ of the clean speech PSD $\sigma_s^2(k,\ell)$ is finally obtained by a re-estimation, again using temporal cepstrum smoothing. The following subsections give a more detailed description of the different components of this proposed single-channel spectral enhancement algorithm.

3.2.1 *Spectral gain estimation*

The gain function used in the spectral enhancement scheme has been proposed in [215] to estimate the spectral magnitude of the clean speech. This estimator is derived by modeling the speech magnitude $|s(k,\ell)|$ as a stochastic variable with a chi probability density function (PDF) with shape parameter μ, while the phase of $s(k,\ell)$ is assumed to be uniformly distributed between $-\pi$ and π. Furthermore, the interference $j(k,\ell) = r(k,\ell) + v(k,\ell)$ is modeled as a complex Gaussian random variable with PSD $\sigma_j^2(k,\ell)$. Assuming that $r(k,\ell)$ and $v(k,\ell)$ are uncorrelated, $\sigma_j^2(k,\ell)$ can be expressed as

$$\sigma_j^2(k,\ell) = \mathrm{E}\left\{|j(k,\ell)|^2\right\} = \sigma_v^2(k,\ell) + \sigma_r^2(k,\ell). \tag{3.18}$$

The squared distance between the amplitudes (to the power β) of the clean speech $s(k,\ell)$ and the estimated output $\hat{s}(k,\ell)$ is defined as

$$\epsilon(k,\ell) = \left(\left|s(k,\ell)\right|^{\beta} - \left|\hat{s}(k,\ell)\right|^{\beta}\right)^{2}, \tag{3.19}$$

where the parameter β, typically chosen as $0 < \beta \leq 1$, denotes a compression factor. This compression factor results in a different emphasis given on estimation errors for small amplitudes in relation to large amplitudes. The clean speech magnitude is estimated by optimizing the MMSE criterion

$$|\hat{s}(k,\ell)| = \underset{|\hat{s}(k,\ell)|}{\operatorname{argmin}} \ \mathrm{E}\left\{\epsilon(k,\ell) | \tilde{y}(k,\ell), \sigma_{\jmath}^{2}(k,\ell), \xi(k,\ell)\right\}, \tag{3.20}$$

where $\xi(k,\ell)$ denotes the *a priori* signal-to-interference ratio (SIR) defined as

$$\xi(k,\ell) = \frac{\sigma_{s}^{2}(k,\ell)}{\sigma_{r}^{2}(k,\ell) + \sigma_{v}^{2}(k,\ell)}, \tag{3.21}$$

where $\sigma_{s}^{2}(k,\ell)$ denotes the PSD of the clean speech.
As shown in [215], the solution to (3.20) leads to the spectral gain $\tilde{g}(k,\ell)$

$$\begin{aligned} \tilde{g}(k,\ell) = {} & \sqrt{\frac{\xi(k,\ell)}{\mu + \xi(k,\ell)}} \cdot \\ & \left[\frac{\mathrm{Gam}\left(\mu + \frac{\beta}{2}\right)}{\mathrm{Gam}\left(\mu\right)} \frac{\Phi\left(1 - \mu - \frac{\beta}{2}, 1; -\nu(k,\ell)\right)}{\Phi\left(1 - \mu, 1; -\nu(k,\ell)\right)}\right]^{1/\beta} \cdot \\ & \left(\sqrt{\gamma(k,\ell)}\right)^{-1}, \end{aligned} \tag{3.22}$$

where $\gamma(k,\ell)$ denotes the *a posteriori* SIR, defined as

$$\gamma(k,\ell) = \frac{|\tilde{y}(k,\ell)|^{2}}{\sigma_{r}^{2}(k,\ell) + \sigma_{v}^{2}(k,\ell)}, \tag{3.23}$$

and

$$\nu(k,\ell) = \frac{\gamma(k,\ell)\xi(k,\ell)}{\mu + \xi(k,\ell)}, \tag{3.24}$$

where $\Phi(\cdot)$ denotes the confluent hypergeometric function and $\mathrm{Gam}\,(\cdot)$ denotes the complete Gamma function [220]. Depending on the choice of β and μ, the solution in (3.22) can resemble other well-known estimators, such as the short-time spectral amplitude estimator ($\beta = 1$, $\mu = 1$) [98] or approaches the log-spectral amplitude estimator [221] when β tends to 0. In order to reduce artefacts that may be introduced by directly applying (3.22), the spectral gain $g(k,\ell)$ in (3.17) is restricted to values larger than a spectral floor $g_{\min}$ (see Section 3.3), i.e.,

$$g(k,\ell) = \max\left(\tilde{g}(k,\ell), g_{\min}\right). \tag{3.25}$$

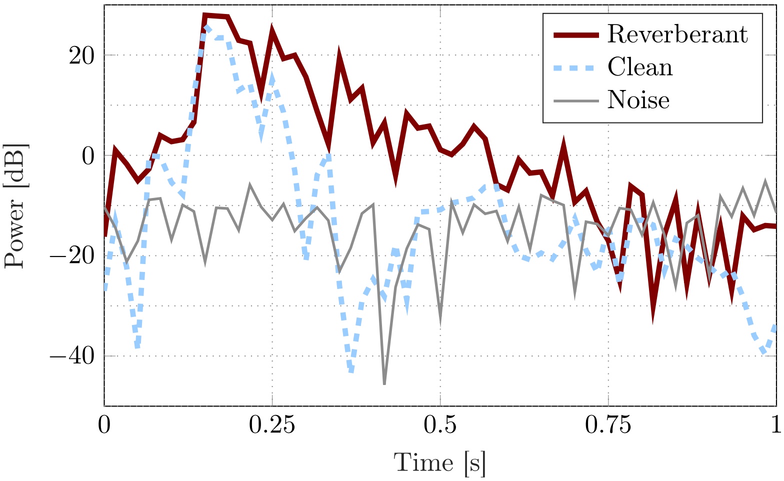

Fig. 3.3: Power of noise as well as of clean and reverberant speech for one second of signal.

To compute the expression in (3.22), the PSDs $\sigma_s^2(k,\ell)$, $\sigma_v^2(k,\ell)$ and $\sigma_x^2(k,\ell)$ have to be estimated from the beamformer output. The used estimators are described in the next subsections.

3.2.2 *Noise PSD estimation*

The MS [90] approach has been shown to be a reliable estimator of the noise PSD for moderately time-varying noise conditions. This approach relies on the assumption that the minimum of the noisy speech power over a short temporal sliding window is not affected by the speech. The noise PSD $\sigma_v^2(k,\ell)$ is then estimated by tracking the minimum of the noisy speech power over this sliding window, whose usual length corresponds to 1.5 s according to [90].

Fig. 3.3 depicts the powers of anechoic speech, reverberant speech and additive noise for one frequency bin of their power spectrograms. As illustrated in this figure, the decay time in speech pauses is typically increased in the presence of reverberation. Consequently, a longer tracking window is used in the proposed spectral enhancement scheme (see Section 3.3) in order to avoid reverberant speech affecting the estimation of the noise PSD $\sigma_v^2(k,\ell)$.

3.2.3 *Speech PSD estimation*

Temporal cepstrum smoothing, as proposed in [216], is used to estimate the PSD $\sigma_x^2(k,\ell)$ of the reverberant speech component $x(k,\ell)$ as well as the PSD $\sigma_s^2(k,\ell)$ of the dereverberated speech signal $s(k,\ell)$. The estimation of $\sigma_x^2(k,\ell)$ only requires the noise PSD estimate $\hat{\sigma}_v^2\,(k,\ell)$ whereas the estimation of $\sigma_s^2(k,\ell)$ additionally requires the estimate $\hat{\sigma}_x^2(k,\ell)$ of the reverberant speech PSD, as depicted in Fig. 3.2. The modifications required for the latter case are described at the end of this section.

The estimate of the reverberant speech PSD $\sigma_x^2(k,\ell)$ is computed using the maximum likelihood (ML) estimator of the *a priori* SNR,

$$\xi_{x_{\text{ml}}}(k,\ell) = \frac{|\tilde{y}(k,\ell)|^2}{\sigma_v^2(k,\ell)} - 1. \tag{3.26}$$

An estimate $\hat{\sigma}^2_{x_{\text{ml}}}(k,\ell)$ of the reverberant speech PSD can then be obtained as

$$\hat{\sigma}^2_{x_{\text{ml}}}(k,\ell) = \hat{\sigma}_v^2(k,\ell)\ \max\left(\xi_{x_{\text{ml}}}(k,\ell),\ \xi_{\text{ml}}^{\min}\right), \tag{3.27}$$

where $\xi_{\text{ml}}^{\min} > 0$ denotes a lower bound set to avoid negative or very small values of $\xi_{x_{\text{ml}}}(k,\ell)$. In the cepstral domain, $\hat{\sigma}^2_{x_{\text{ml}}}(k,\ell)$ can be represented by

$$\lambda_{x_{\text{ml}}}(q,\ell) = \text{IFFT}\left\{\log\left(\hat{\sigma}^2_{x_{\text{ml}}}(k,\ell)|_{k=0,\cdots,(L-1)}\right)\right\}, \tag{3.28}$$

where q and L denote the cepstral bin index and the length of the FFT, respectively. A recursive temporal smoothing is applied to $\lambda_{x_{\text{ml}}}(q,\ell)$, i.e.,

$$\lambda_x(q,\ell) = \delta(q,\ell)\lambda_x(q,\ell-1) + (1-\delta(q,\ell))\lambda_{x_{\text{ml}}}(q,\ell), \tag{3.29}$$

where $\delta(q,\ell)$ denotes a time-quefrency-dependent smoothing parameter. Only a mild smoothing is applied to the quefrencies which are mainly related to speech, while for the remaining quefrencies a stronger smoothing is applied. Consequently, a small smoothing parameter is chosen for the low quefrencies, as they contain information about the vocal tract shape, and for the quefrencies corresponding to the fundamental frequency f_0 in voiced speech. In order to protect these quefrencies, especially the ones corresponding to the fundamental frequency, the parameter $\delta(q,\ell)$ in (3.29) is adapted. After determining f_0 by picking the highest peak in the cepstrum within a limited search range, $\delta(q,\ell)$ is defined as

$$\delta(q,\ell) = \begin{cases} \delta_{\text{pitch}} & \text{if } q \in \mathbb{Q}, \\ \bar{\delta}(q,\ell) & \text{if } q \in \{0,\cdots,L/2\} \setminus \mathbb{Q}, \end{cases} \tag{3.30}$$

where $\mathbb{Q}$ denotes a small set of cepstral bins around the quefrency corresponding to f_0 and where δ_{pitch} denotes the smoothing parameter for the quefrency bins within $\mathbb{Q}$ [216]. The quantity $\bar{\delta}(q,\ell)$ is given as

$$\bar{\delta}(q,\ell) = \eta\delta(q,\ell-1) + (1-\eta)\bar{\delta}_{\text{const}}(q), \tag{3.31}$$

where $\bar{\delta}_{\text{const}}(q)$ is time-independent and chosen such that less smoothing is applied in the lower cepstral bins. Furthermore, η denotes a forgetting factor that defines how fast the transition from $\delta(q,\ell)$ to $\bar{\delta}_{\text{const}}(q)$ can occur (see Section 3.3). Finally, the reverberant speech PSD estimate $\hat{\sigma}_x^2(k,\ell)$ can be computed by transforming $\lambda_x(q,\ell)$ back to the spectral domain, i.e.

$$\hat{\sigma}_x^2(k,\ell) = \exp\left(\kappa + \mathrm{DFT}\left\{\lambda_x(q,\ell)\right\}|_{q=0,\cdots,(L-1)}\right), \tag{3.32}$$

where κ denotes a parameter used to compensate for the bias due to the recursive smoothing in the log-domain in (3.29) and is estimated as in [217].

The estimate of the reverberant speech PSD can be used to estimate the (late) reverberant PSD $\sigma_r^2(k,\ell)$ (see Subsection 3.2.4). After having estimated $\sigma_r^2(k,\ell)$, cepstral smoothing is also used to estimate the dereverberated clean speech PSD $\sigma_s^2(k,\ell)$. In this case, the noise PSD $\sigma_v^2(k,\ell)$ in equations (3.26) and (3.27) is replaced by the interference PSD $\sigma_j^2(k,\ell) = \sigma_v^2(k,\ell) + \sigma_r^2(k,\ell)$.

3.2.4 *Reverberant PSD estimation*

The RIR model presented in [24] represents the RIR as a Gaussian noise signal multiplied by an exponential decay Δ, which depends on the room reverberation time, T_{60}, i.e.,

$$\Delta = \frac{3\ln(10)}{\mathrm{T}_{60} f_s}. \tag{3.33}$$

In the proposed spectral enhancement algorithm, the approach derived from this model and presented in [87] is used to estimate the reverberation PSD $\sigma_r^2(k,\ell)$ as

$$\hat{\sigma}_r^2(k,\ell) = \mathrm{e}^{-2\Delta T_d f_s}\hat{\sigma}_x^2(k,\ell - T_d/T_s), \tag{3.34}$$

where

$$\hat{\sigma}_x^2(k,\ell) = \hat{\sigma}_r^2(k,\ell) + \hat{\sigma}_s^2(k,\ell). \tag{3.35}$$

In (3.34), T_s denotes the frame shift whereas T_d is the duration of the direct path and early reflections of the RIR, typically assumed to be between 50 ms and 80 ms. As a result, the estimate $\hat{\sigma}_r^2(k,\ell)$ of the late reverberation PSD can be obtained using $\hat{\sigma}_x^2(k,\ell)$ and an estimate of the reverberation time T_{60} obtained using an online estimator. A comparison of online T_{60} estimators can be found, e.g., in [79]. In this thesis, as in [195, 196], the T_{60} is estimated using the estimator proposed in [219].

Finally, using the estimated PSDs of the reverberation and of the residual noise, an estimate $\hat{\sigma}_s^2(k,\ell)$ of the clean speech PSD is obtained. These estimates are used in (3.21) to compute the *a priori* SIR and in (3.22) to compute the real-valued spectral gain, $\tilde{g}(k,\ell)$.

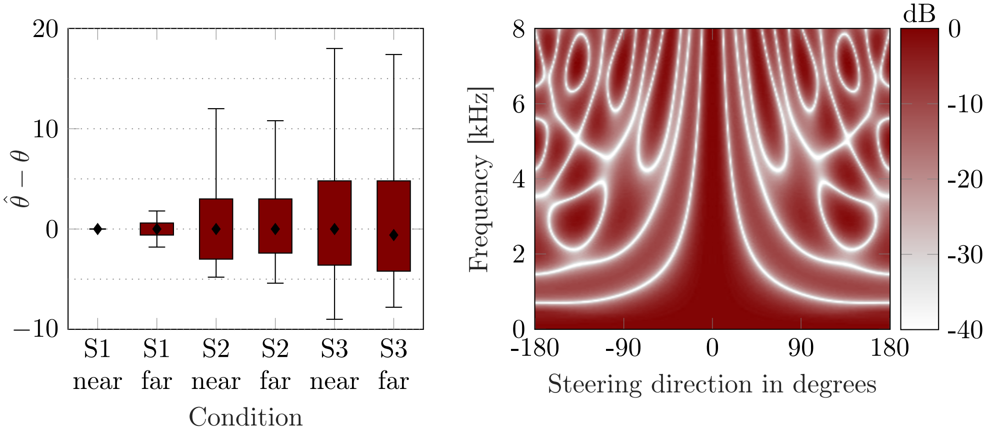

Fig. 3.4: Error in DOA estimation obtained on the simulated data of the REVERB challenge corpus (left) and beampattern of the used MVDR beamformer computed using the noise coherence matrix of a theoretically diffuse noise field (right).

3.3 Experimental setup

3.3.1 *Algorithm settings*

The experiments are conducted using the corpus described in Chapter 2 in which all signals are recorded at a sampling frequency $fs = 16\,\text{kHz}$. As both the T_{60} and the DOA of the target speaker are assummed constant, both T_{60} and DOA have been estimated only once per utterance as their median estimate over all frames. The STFT has been computed using a 32 ms Hann window with 50 % overlap and an FFT of length $L = 512$. The DOA has been estimated as the angle minimizing the sum of the MUSIC pseudo-spectra, for $\theta = 0° \ \ldots \ 360°$ for every 2°, using all 8 microphones of the circular microphone array for the frequency range from 50 Hz to 5 kHz, see Subsection 3.1.3.

The MVDR beamformer uses a theoretically diffuse noise coherence matrix and a white noise gain constraint WNG_{max} = -10 dB if less than 10 frames are detected as noise when applying the VAD, see (3.11). The VAD has been configured similarly as in [89] but its parameters have been adapted in order to apply it to signals with a sampling frequency of 16 kHz. Otherwise, the noise coherence matrix is estimated using all detected noise-only frames, see (3.9). The speech amplitude estimator described in Subsection 3.2.1 assumes a chi PDF with shape parameter $\mu = 0.5$, a minimum gain g_{min} of -10 dB and a compression parameter $\beta = 0.5$. The noise PSD estimator described in Subsection 3.2.2 uses the same parameters as in [90], except for the length of the sliding window for minima tracking which has been set to either 1.5 s or 3 s for the configurations that in our experiments are denoted by $\text{SE}_{1.5}$ and SE_3, respectively. In (3.31), $\eta = 0.96$ and all parameters used used for the speech PSD estimation, described in Subsection 3.2.3, have been set as prescribed in [215]. In equation (3.34), T_d has been set to 80 ms.

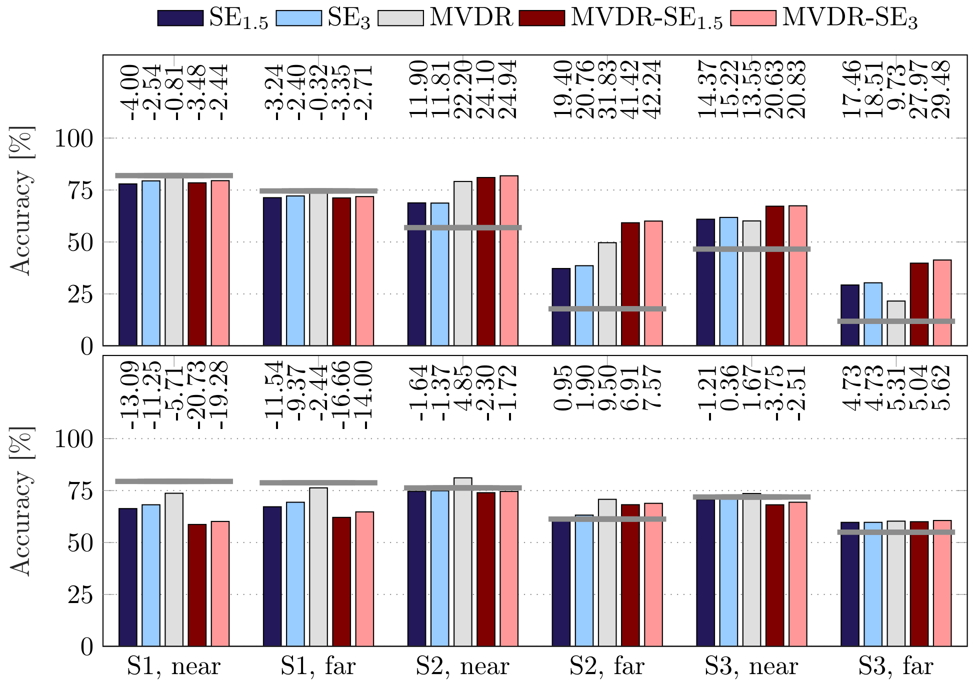

Fig. 3.5: Speech recognition accuracy on the simulated data of the considered corpus when using an acoustic model trained using either clean data (top) or multi-condition data (bottom). Horizontal bars mark the accuracy obtained on unprocessed data while numbers indicate the accuracy gain obtained by applying the considered algorithms.

3.3.2 *Observations on beamformer design*

The MVDR beamformer is steered towards the estimated DOA of the target speech signal. In practice, errors in the DOA estimation can result in speech degradation. Fig. 3.4 (left) depicts the DOA error obtained in all conditions for the simulated data of the REVERB challenge (i.e. a total of 2176 utterances). The true DOA has been considered to be the one stated in the REVERB challenge data documentation [203]. Ignoring outliers, it can be seen that the absolute value of the error is smaller than 5 degrees in room S1 while in room S2, it is smaller than 5 degrees for 50% of the data and always smaller than 15 degrees. As expected, the largest error in DOA estimation appears in the case of room S3, which has the largest reverberation time. It can be seen that for room S3, for 50% of the utterances, the absolute value of the DOA error is inferior to 5 degrees. However, it can be close to 20 degrees for some utterances.

In order to assess the detrimental effect that such DOA error could have on the performance of the MVDR beamformer, one may examine its corresponding beampattern. Fig. 3.4 (right) depicts the beampattern of the MVDR beamformer computed

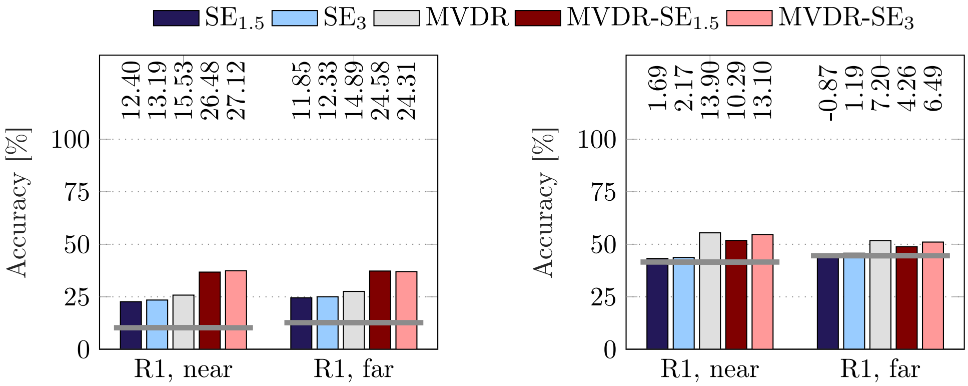

Fig. 3.6: Speech recognition accuracy on the real data of the considered corpus when using an acoustic model trained using either clean data (left) or multi-condition data (right). Horizontal bars mark the accuracy obtained on unprocessed data while numbers indicate the accuracy gain obtained by applying the considered algorithms.

using the noise coherence matrix of a theoretically diffuse noise field as in (3.11), steered towards the zero degrees direction, and using the microphone configuration described in Section 2.2. By observing the width of the main lobe it appears that the error in DOA is small enough to not introduce distortions in room S1 and S2. Some cancellation of the target speech signal may occur in room S3 but should be limited to frequencies higher than 4 kHz.

3.4 Results

3.4.1 *Speech recognition performance*

In order to evaluate the potential benefit of the proposed signal enhancement algorithm on the performance of an ASR system, the processed signals have been used as the input for the baseline ASR system provided by the REVERB challenge (see Section 2.2). The ASR accuracy is obtained by extracting the MFCCs, including Deltas and double-Deltas, from the processed signal $\hat{s}(n)$ for both simulated and real data. These MFCCs are then input to the ASR system. This subsection presents the results obtained when using acoustic models trained on either clean or multi-condition training data.

Fig. 3.5 depicts the accuracy obtained on the simulated data using either clean (top) or multi-condition training data (bottom). When using the model trained on clean data and comparing the accuracy to the one obtained from unprocessed signals (see horizontal lines in Fig. 3.5), the accuracy decreases slightly for the conditions with the lowest reverberation time (room S1). This indicates that spectral coloration introduced by the enhancement algorithm may reduce the performance of the ASR system while the benefit of dereverberation is limited for small reverberation times.

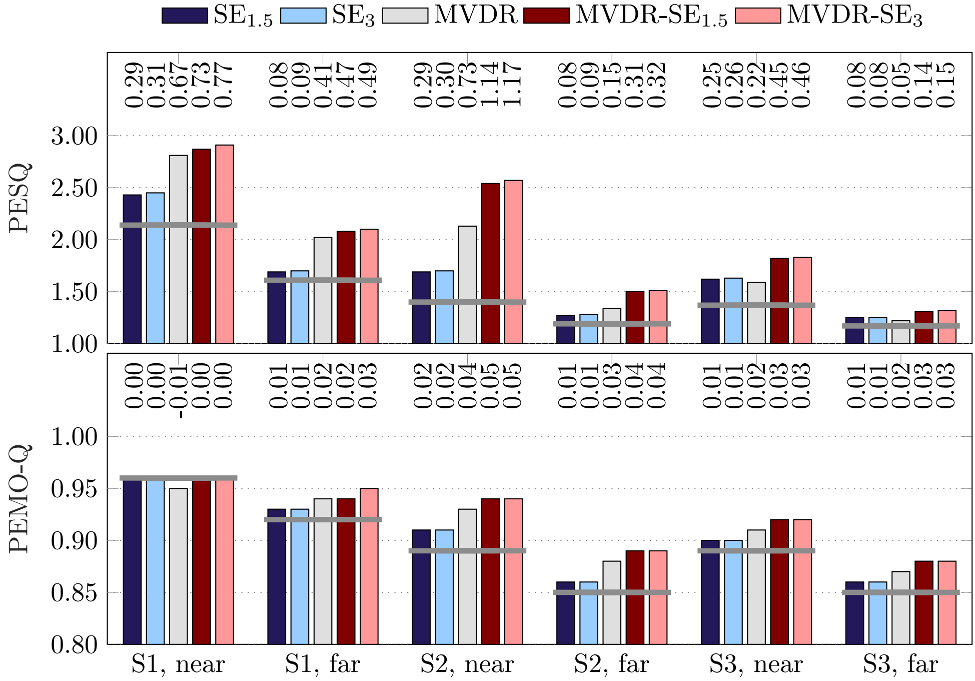

Fig. 3.7: PESQ (top) and PEMO-Q (bottom) scores computed on the simulated data of the REVERB challenge corpus. Horizontal bars mark the scores obtained on unprocessed data while numbers indicate the performance gain obtained by applying the considered algorithms.

In all other conditions, the single-channel spectral enhancement algorithm improves the accuracy, with SE_3 yielding larger improvements than $SE_{1.5}$. Except for room S3, the MVDR beamformer yields better results than the single-channel algorithm. The combination of the MVDR beamformer with SE_3 yields the largest improvement: absolute accuracy improvement up to 42.24 % for the simulated data (condition “S2, far”).

On unprocessed signals from room S1, it appears that using the multi-condition model does not improve the accuracy obtained with the clean model. In addition, it appears that in room S1 all considered algorithms reduce the accuracy with this detrimental effect being the least noticeable when applying MVDR alone. This degradation of performance might be due to the inability of the considered ASR system to deal with distortions, even minors, that can be introduced by speech enhancement algorithm. Indeed, the training data did not entails processed signal and the MVDR, which except in the case of DOA estimation error would be distortionless, would logically be less affected. However, for rooms S2 and S3, multi-condition training does improve accuracy and there is not a severe drop in performance when the acoustic conditions are more challenging. A notable effect to observe is the fact that, using multi-condition training, the accuracy improvement obtained by applying the considered algorithms is quite minor. Moreover, the difference in performance between the considered algorithms is very small. Notably, for condition "S3, far", the gain in performance between the $\mathrm{SE}_{1.5}$ and MVDR-SE_3 is of less than a percentage point.

The performance obtained on real data is depicted in Fig. 3.6. When using the acoustic model trained on clean data, all considered algorithms improve accuracy with the largest improvement coming from the combination of beamforming and spectral suppression. More specifically, the highest improvement is obtained by MVDR-SE_3 in condition "R1, near", with an accuracy improvement of 27.12 percentage points. By observing the accuracy obtained by using the acoustic model trained on multi-condition data, it appears that in this more challenging condition, this model is advantageous. However, similarly as what was observed from the simulated data, the difference in performance between the considered algorithms is limited. It appears that for ASR accuracy on the considered benchmark, MVDR is the most suited algorithm. However, the best choice of algorithms for ASR might not be the same as for speech quality improvement, whose evaluation through signal-based measures and listening tests is presented in the next subsections.

3.4.2 *Signal-based measures*

This section evaluates the performance of the considered speech enhancement algorithms by applying the signal based measures listed in Chapter 2. The performance in term of PESQ and PEMO-Q is depicted in Fig. 3.7. As both PESQ and PEMO-Q are intrusive measures, they are only provided for the simulated data. Meanwhile the considered non-intrusive measures, namely ANIQUE+, P.563 and $\mathrm{SRMR}_{\mathrm{norm}}$, are presented for both simulated and real data in Fig. 3.8 and Fig. 3.9, respectively. As expected, the signal based-measures do not always show completely consistent results. Nevertheless, some common tendencies can clearly be observed and are summarized in the following.

Both PESQ and PEMO-Q aim at estimating the overall speech quality of the test signal intrusively. The PESQ and PEMO-Q scores are depicted in Fig. 3.7. PESQ shows the same trend in all conditions. More specifically, all considered enhancement algorithms yield higher scores than the unprocessed signal and multichannel enhancement yields the largest improvement. A minor advantage of MVDR+SE_3 over MVDR+$\mathrm{SE}_{1.5}$ appears, e.g., in condition "S1, far". PEMO-Q shows similar trends, except for condition "S1, near" in which it does not show any difference between processed and unprocessed speech, and even yields a slightly lower score in the case of MVDR. One cannot reach definitive conclusions about the benefits of the considered algorithms on the basis of PESQ and PEMO-Q alone and should examine other signal-based measures.

The scores of the non-intrusive measures, i.e., ANIQUE+, P.563 and $\mathrm{SRMR}_{\mathrm{norm}}$ are depicted in Fig. 3.8. As was the case for the intrusive measures, all indicate that the considered algorithms improve the speech quality compared to their unprocessed counterpart and that multichannel enhancement yields the largest improvement. All three considered non-intrusive measures show MVDR as providing nearly no im-

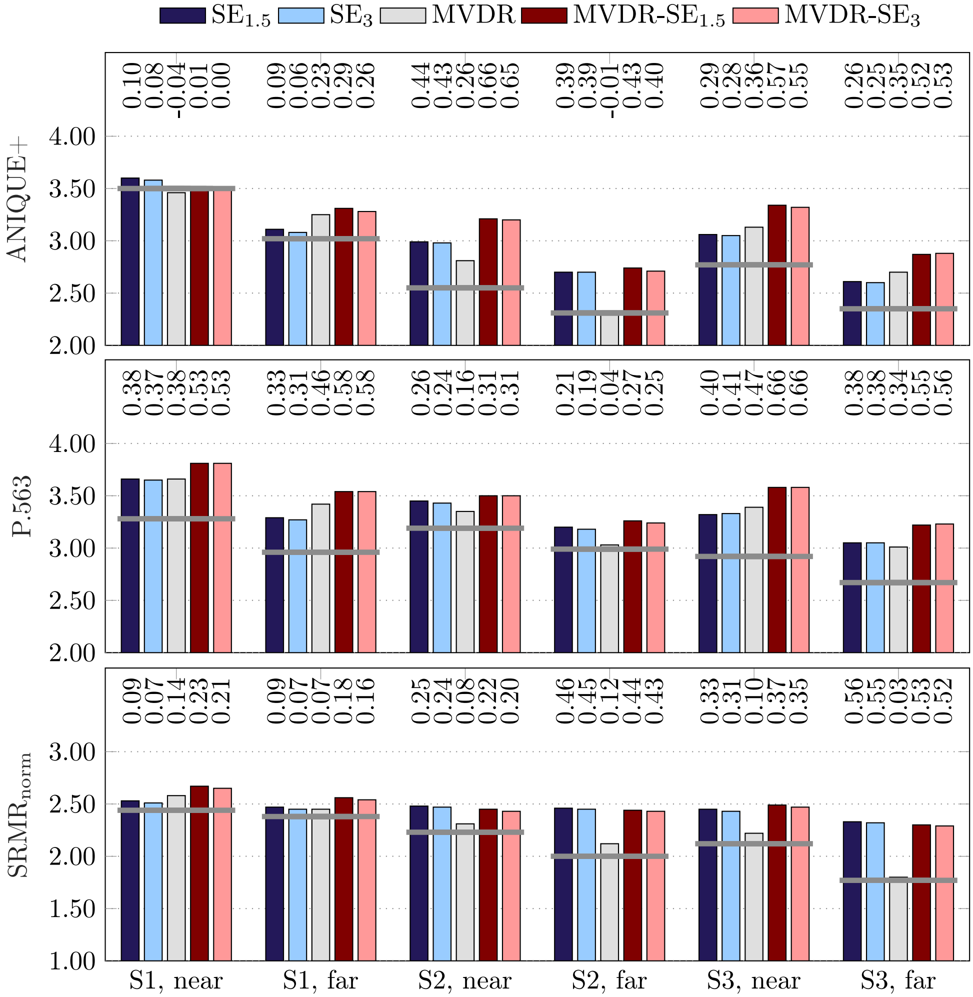

Fig. 3.8: ANIQUE+ (top), P.563 (middle) and $\text{SRMR}_{\text{norm}}$ (bottom) scores computed using the simulated data of the REVERB challenge corpus. Horizontal bars mark the scores obtained on unprocessed data while numbers indicate the performance gain obtained by applying the considered algorithms.

provement for condition “S2, far”, though the combined algorithms MVDR+$\text{SE}_{1.5}$ and MVDR+SE_3 appear to yield better performance than their single-channel counterparts. This performance of MVDR is an example where the measures show inconsistency between one another, e.g., it appears beneficial when looking at ANIQUE+ or P.563 but seems to yield the same quality as unprocessed signals when considereing $\text{SRMR}_{\text{norm}}$ alone. In addition, the minor advantage that MVDR+SE_3 seems to have over MVDR+$\text{SE}_{1.5}$ when considering the intrusive measure does not appear with these three non-intrusive measures.

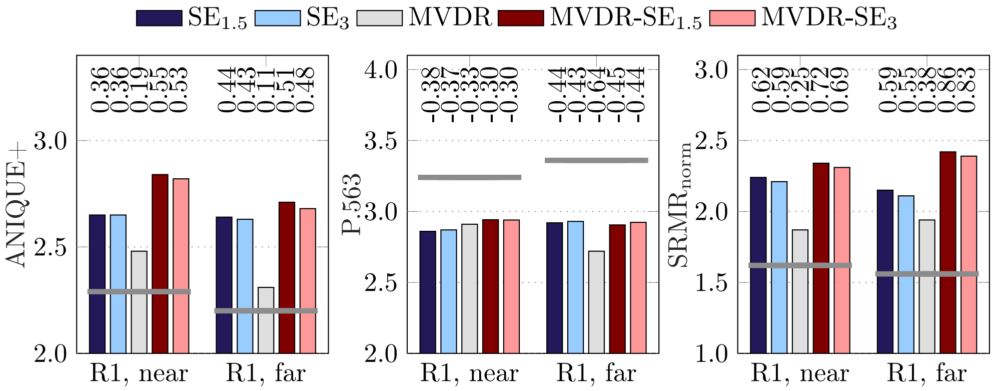

Fig. 3.9: ANIQUE+ (left), P.563 (middle) and $\mathrm{SRMR}_{\mathrm{norm}}$ (right) scores computed using the real data of the REVERB challenge corpus. Horizontal bars mark the scores obtained on unprocessed data while numbers indicate the performance gain obtained by applying the considered algorithms.

An even larger inconsistency between the measures appears when observing the scores obtained on the real data depicted in Fig. 3.9. Both ANIQUE+ and $\mathrm{SRMR}_{\mathrm{norm}}$ indicate that all algorithms improve the speech quality, that MVDR is the least beneficial algorithm and that the combined algorithms MVDR+$\mathrm{SE}_{1.5}$ and MVDR+SE_3 yield better performance than $\mathrm{SE}_{1.5}$ and SE_3. However, P.563 indicates that all algorithms are detrimental when applied to real data, which is in strong contradiction to the other measures.

By observing all considered signal-based measures, it seems that the proposed algorithms are beneficial to the speech quality and that, as could be expected, MVDR+SE_3 is the most advantageous. However, due to the lack of consistency between the measures, perceptual evaluation, i.e., listening tests, is required to reliably assess the performance of these algorithms.

3.4.3 *Perceptual evaluation*

This subsection presents the results of the perceptual evaluation that we conducted and first reported in [196] and reproduces the results of the online perceptual evaluation conducted by the organisers of the REVERB challenge and reported in [31]. These subjective evaluations are based on MUSHRA tests following the specifications described in [37]. Four acoustic conditions have been tested, “S2, near”, “S2, far”, “R1, near” and “R1, far”, as in the work presented in [31].

3.4.3.1 *Perceptual evaluation in controlled environment*

This perceptual evaluation has been carried out for the unprocessed signals and for 3 processing algorithms, namely, the single-channel scheme applied to the first microphone signal (SE_3), the MVDR beamformer using 8 microphones (MVDR) and

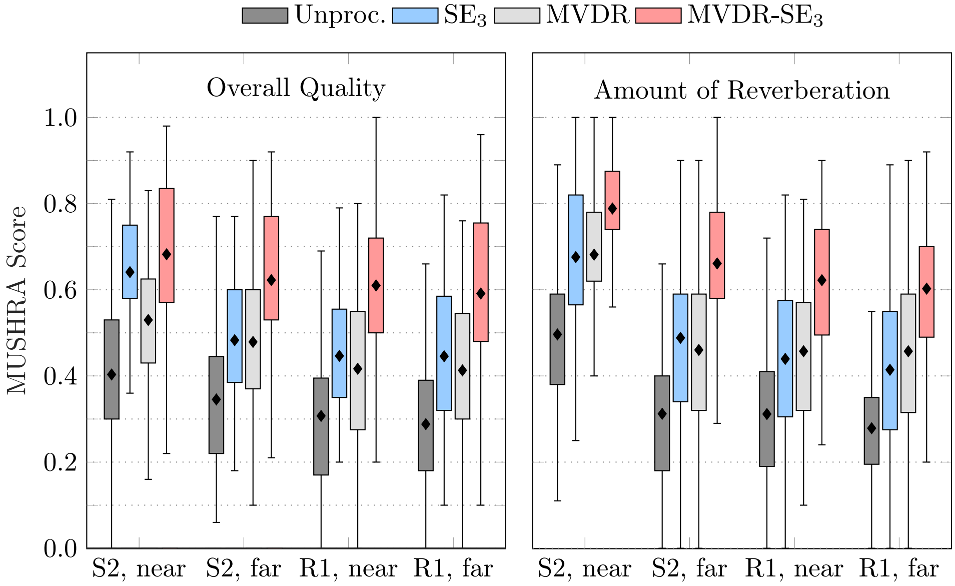

Fig. 3.10: MUSHRA scores from our perceptual evaluation, normalized to range between 0 and 1. Higher scores indicate a higher quality (left) or a larger reverberation suppression (right), diamonds indicate the mean. Scores for the reference, always labeled as excellent, and of the anchor, assessed as least satisfactory, are not displayed.

the combination of the MVDR beamformer with the spectral enhancement scheme (MVDR+SE$_3$). In addition to these signals, a hidden reference and an anchor have been presented to the assessors. The hidden reference was the anechoic speech signal in the case of simulated data and the signal recorded by a headset microphone in the case of real data. The anchor consisted of the first microphone signal, low-pass filtered with a cut-off frequency of 3.5 kHz.

A total of 21 self-reported normal-hearing listeners participated in the MUSHRA listening test. The listening test was conducted in a soundproof booth and the assessors listened to diotic signals through headphones (Seinheiser HD 380 pro). Each assessor evaluated 3 utterances per condition (i.e. 12 utterances per assessor), in terms of two different attributes: “overall quality” and “perceived amount of reverberation”, on a scale ranging from 0 to 100. For each assessor, the utterances to be evaluated were randomly picked from the REVERB challenge database. All signals were normalized in amplitude and presented at a sampling frequency of 16 kHz and a quantization of 16 bit using a Roland sound card (model UA-25EXCW). The listening test was divided into three stages.

In the first stage, the assessors were asked to listen to all files that would be presented to them during a training phase. This training phase allowed the assessors to get familiar with the data to be evaluated and to adjust the sound volume to a comfortable level. In the second stage, the assessors had to evaluate the overall quality

Table 3.1: Results of the Friedman's test for both tested attributes. The value $p < 0.01$ indicates the significance of the results and χ^2 denotes the Friedman's chi square statistic.

		S2, near	S2, far	R1, near	R1, far
Overall quality	χ^2	99.6	77.1	98.9	90.6
	p	<0.01	<0.01	<0.01	<0.01
Amount of reverberation	χ^2	93.9	120.8	98.6	104.7
	p	<0.01	<0.01	<0.01	<0.01

of the signals. Finally, the third stage consisted in the evaluation of the perceived amount of reverberation. The order of presentation of algorithms and conditions were randomized between all stages and all assessors.

The obtained MUSHRA scores, normalized to range between 0 and 1, are summarized in Fig. 3.10. The anchor appears to be the least satisfactory for the attribute "overall quality", suggesting that the subjects used the full extent of the grading scale. However, this is not the case for the attribute "perceived amount of reverberation", illustrating the difficulty of evaluating this attribute. The three considered processing algorithms yielded an improvement compared to the unprocessed signal both in terms of "overall quality" and of "perceived amount of reverberation". As expected, the largest reduction of the "perceived amount of reverberation" is observed for the combination MVDR+SE_3. The combination MVDR+SE_3 improves the overall quality as well, although the improvement, compared to the single-channel algorithm, is lower than for the attribute "perceived amount of reverberation". The use of an MVDR beamformer alone reduces the "perceived amount of reverberation" but does not improve the performance compared to SE_3.

Since the scores of the MUSHRA test were not normally distributed, a Friedman's test [222] was used to examine the significance of the results, excluding the scores of the anchor and the reference. The results of the Friedman's test are presented in Table 3.1. The p-value, $p < 0.01$, shows that at least one significant pairwise difference can be observed in all conditions and for all attributes. In order to examine the significance of the pairwise difference in performance between the algorithms, a Wilcoxon rank sum test [223] has been used for each condition separately. A Bonferroni correction has been applied resulting in significant effects being considered for $p < 0.05/6$. For the attribute "perceived amount of reverberation", the differences in performance between the unprocessed signal and all algorithms are significant but no significant differences were present between the different algorithms. The same conclusion holds for the attribute "overall quality", except for the room R1 and the condition "S2, near", where the differences between the unprocessed signal and the output of the MVDR beamformer do not appear to be significant.

Even though the statistical significance criterion is not always satisfied, the trend of the results confirms the benefits of combining a beamformer with a single-channel spectral enhancement algorithm for reducing reverberation and noise and for improving the overall speech quality.

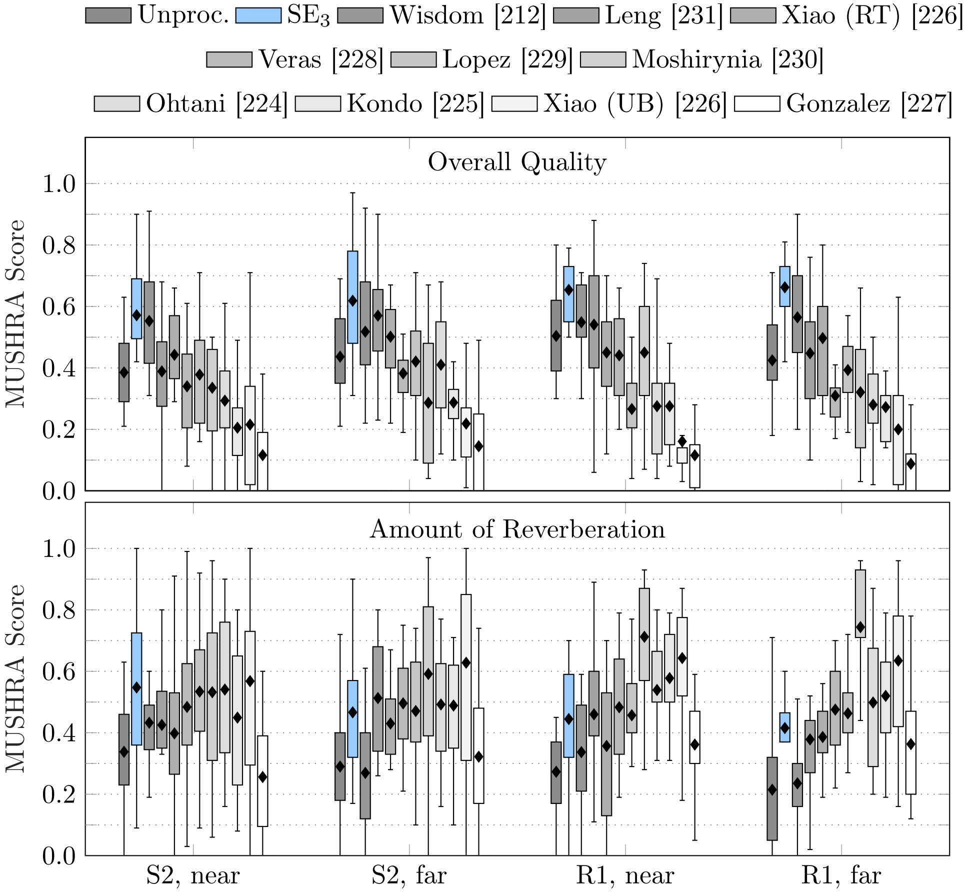

Fig. 3.11: MUSHRA scores for single-channel algorithms obtained during the test conducted online by the REVERB challenge organizers [31]. Higher scores indicate a higher quality (top) or a larger reverberation suppression (bottom), diamonds indicate the mean. Reference and anchor are excluded and scores are normalized between 0 and 1.

3.4.3.2 *Online perceptual evaluation*

A large scale listening test was conducted as part of the REVERB challenge to measure, as in the previously describe perceptual evaluation in controlled environment, the overall speech quality and level of reverberation. As both SE$_3$ and MVDR+SE$_3$ were submitted to this challenge, we present below the results of this listening test that were published in [31], for which the authors kindly provided the raw scores and where a more detailed test description can be found. This test was conducted using crowdsourcing over internet. Compared to perceptual evaluation in controlled environment, this method has both drawbacks and advantages. Indeed, though assessors were instructed to conduct the test over headphones and in a quiet environment,

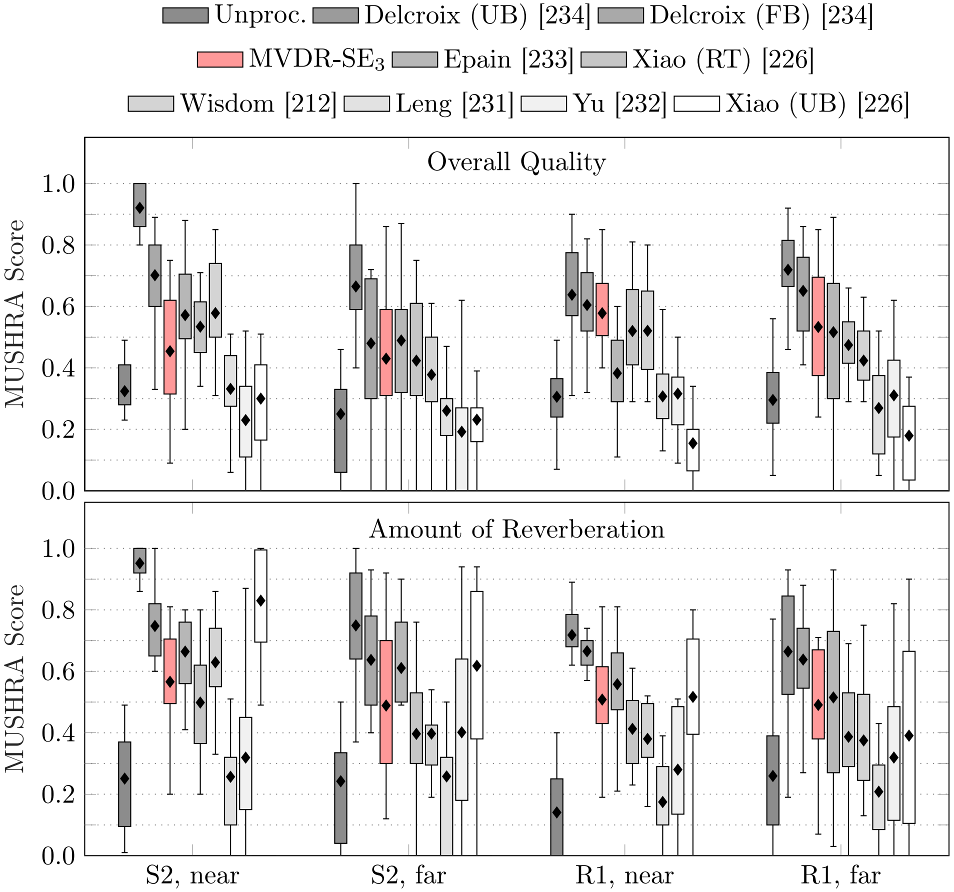

Fig. 3.12: MUSHRA scores for multichannel algorithms obtained during the test conducted online by the REVERB challenge organizers [31]. Higher scores indicate a higher quality (top) or a larger reverberation suppression (bottom), diamonds indicate the mean. Reference and anchor are excluded and scores are normalized between 0 and 1.

it was not possible to control that they followed the instructions adequately. Furthermore, one cannot guarantee that the audio system of the assessors, e.g., used sound card, did not introduce artefacts. However, this test allowed for the comparison of a wide range of algorithms by a wide range of participants, which would not have been feasible in a controlled environment. Though these results should only be interpreted with caution, some important observations can be made.

In this online perceptual evaluation, single-channel and multichannel algorithms were evaluated separately. Consequently, only signals processed from the same type of input, e.g., single-channel, and in the same condition, e.g., “S2, near”, were presented simultaneously to assessors. In this thesis, the MUSHRA scores from this evaluation, normalized between 0 and 1, are presented for the single-channel algorithms (including SE_3) in Fig. 3.11 and for the multichannel algorithms (including MVDR+SE_3) in Fig. 3.12. In both figures, the algorithms are denoted by the name of the first author and reference number of the REVERB challenge paper in which they were presented, i.e, for single-channel algorithms, [212, 224–231], and for multichannel algorithms, [212, 226, 231–234]. The acronyms in brackets are provided for algorithms that have been applied in several settings that can be real time (RT), utterance-based (UB) or full batch (FB) processing [31]. These indicate that the estimators used in these algorithms have been applied either in real time, using complete utterances or the complete dataset.

By observing the scores depicted in Fig. 3.11, it appears that many single-channel algorithms are able to reduce the perceived amount of reverberation but that few are able to do so while improving the perceived speech quality. However, the proposed SE_3 algorithm was able to do both and outperformed all other single-channel algorithms in terms of perceived speech quality. This shows that SE_3 is a robust baseline to use for future developments of speech enhancement algorithms. In the multichannel case, for which scores are presented in Fig. 3.12, the combined algorithm MVDR+SE_3 is as well able to both reduce reverberation and improve speech quality. Though, using the 8 available channels, MVDR+SE_3 is not the best of the considered algorithms, it is only outperformed by the algorithm from Delcroix et al. that relies on a linear prediction step that can be computationally intensive. Consequently, MVDR+SE_3 is a valuable algorithm to be used when multiple microphones are available but computing power is limited.

3.5 Summary

In this chapter, we have presented the combination of an MVDR beamformer with a single-channel spectral enhancement algorithm, aiming at joint dereverberation and noise reduction. In the MVDR beamformer the noise coherence matrix is estimated online using a VAD, whereas the DOA of the target speaker is estimated using the MUSIC algorithm. The output of this beamformer is processed using a spectral enhancement scheme combining statistical estimators of the speech, noise and reverberant PSDs and aiming at joint residual reverberation and noise suppression. The main contribution of this chapter is the combination of several estimators used in this single-channel spectral enhancement algorithm.

The evaluation of the proposed algorithm was done using both simulated and real signals. This evaluation showed the ability of the proposed algorithm to improve both the performance of an ASR system and the quality of speech in communication applications. Noticeably, the proposed single-channel algorithm is able to reduce the perceived amount of reverberation while improving the perceived speech quality. This appears by examining the results of listening tests as well as the considered signal-based measures. However, this evaluation shows, as well, that signal-based measures might be difficult to interpret. The challenge of designing a reliable non-intrusive signal-based measure of the speech quality is addressed in the next chapter.

4

NON-INTRUSIVE QUALITY PREDICTION FOR PROCESSED SPEECH

Many speech enhancement algorithms have been proposed to overcome the speech degradation that both noise and reverberation often introduce [16, 21, 23, 26]. This degradation can be quite severe, particularly when using one or even several microphones to record the speech of a distant user. Fortunately, many algorithms are able to substantially reduce the amount of noise and reverberation. Among those, the algorithms presented in Chapter 3 were able to reduce degradation, in this case focusing on reverberation, while improving the overall speech quality. However, in many speech communication applications, such as teleconferencing or hearing-aids, the choice of the algorithm best suited to reach this goal often depends on the acoustic condition and processing artefacts may result in a degradation of speech quality [31]. Consequently, speech enhancement algorithms need to be evaluated in terms of speech intelligibility and quality, the latter being the focus of this chapter.

Perceptual speech quality evaluation based on listening tests requires a group of human assessors to evaluate the processed speech signals with respect to predefined attributes, such as overall quality, level of reverberation or residual noise, or coloration. Such evaluation is typically performed by grading each attribute on a scale that consists either of a few values, such as for the mean opinion score (MOS) [235], or of continuous values, as in the MUSHRA test [37], which was used in the previous chapter in both a controlled environment and online. Speech intelligibility can be assessed as the number of speech items, i.e., phonemes or words, identified by assessors in relation to the total number of items present in the signal under test [236]. Speech intelligibility is often reported using the SRT, i.e., the level of degradation for which only 50 % of the speech items are correctly identified by an assessor [155].

This chapter is partly based on:

[197] B. Cauchi, J. Santos, K. Siedenburg, T. Falk, P. Naylor, S. Doclo, and S. Goetze, "Predicting the quality of processed speech by combining modulation-based features and model trees," in *Proc. ITG Conf. on Speech Communication*, Paderborn, Germany, Oct. 2016

[198] B. Cauchi, K. Siedenburg, J. Santos, T. Falk, S. Doclo, and S. Goetze, "Non-intrusive speech quality prediction using modulation energies and LSTM-network," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1151–1163, Jul. 2019

Perceptual measures are generally considered the most reliable way to assess the quality or intelligibility of processed speech signals. However, since these measures are costly and time-consuming, speech enhancement algorithms are often evaluated using signal-based measures, that, as observed in the previous chapter, might be difficult to interpret or might give results that are not meaningful for the task.

Signal-based measures, aiming at either speech quality or speech intelligibility prediction, can be categorized as either intrusive or non-intrusive. The computation of intrusive measures requires a (clean) reference signal in addition to the test signal, whereas non-intrusive measures can be computed from the test signal only. Among intrusive measures, the AI [163], the STI [164], the SII [165], the STOI [166] and mutual-information-based techniques, such as the algorithm proposed in [167], aim at speech intelligibility prediction, whereas the PESQ [176], the POLQA [177] or the PEMO-Q [178] are used to predict the speech quality. However, in practice, a reference signal is not available, e.g., to evaluate algorithms using realistic corpora or to automatically select the best algorithm for a specific acoustic condition. Consequently, reliable non-intrusive measures are required.

Several measures have been proposed to remove the need for a reference signal. Non-intrusive measures of the speech intelligibility include the recently proposed non-intrusive STOI [180], that relies on estimating the amplitude envelope of the clean speech from the input signal, and the use of a trained speech recognizer as proposed in [182, 237]. To evaluate speech quality, non-intrusive measures such as P.563 [184] and ANIQUE+ [192] exist, which have not been explicitly developed for the evaluation of speech enhancement algorithms but rather for the evaluation of narrow-band speech codecs.

Measures such as the SRMR [238] and its extension, $\text{SRMR}_{\text{norm}}$ [186], have been developed for both intelligibility and quality prediction and apply a rather simple predicting function to a set of time-averaged modulation energies. Though these measures have shown promising results, e.g., when using $\text{SRMR}_{\text{norm}}$ to predict intelligibility for cochlear implant users in [187], their performance, similarly as for P.563 and ANIQUE+, can be unpredictable when applied to signals processed with different categories of algorithms, as reported in [188]. In [191], twin HMMs have been proposed to generate an estimate of the clean speech before using this estimate and the signal under test as input to an intrusive measure. The reliability of the obtained prediction largely depends on the accuracy of the estimated clean signal such that the method does not outperform the used intrusive measure.

This chapter describes two non-intrusive measures aiming at reliably predicting the speech quality of processed signals across various acoustic conditions and types of processing. Both measures are motivated by the idea that, using adequate features as input and suitable training data, machine learning techniques would be able to model the perception of speech quality. For this purpose, two different predicting functions are used. First, a model tree that used time-averaged features as input, second, a neural network using long short-term memory (LSTM) cells that takes the time-dependency of the test signal into account. The considered predicting functions are trained on a perceptually evaluated dataset constructed for this purpose.

The remainder of this chapter is structured as follows. In Section 4.1, we present the proposed measures, that combine modulation energy (ME) with either a model tree or a RNN using LSTM cells. Using such network as predicting function allows to model the time-dependency of the test signal and to apply the proposed measure to signals of arbitrary length. In Section 4.2, we describe the perceptually evaluated dataset of speech signals, representing various acoustic conditions, i.e., RIRs, noise types and SNRs, and several categories of algorithms, single- and multichannel, with different settings resulting in various level of interference suppression and processing artefacts. Section 4.3 describes how this dataset is used to train and evaluate the proposed measures and presents the experimental framework. The results presented in Section 4.4 show that the LSTM based measure is the most promising and that, when trained for a single category of algorithms, it outperforms existing non-intrusive measures and yields similar performance as the intrusive measures. When considering several categories of algorithms, the LSTM based measure outperforms both non-intrusive and intrusive measures included in the considered benchmark.

4.1 Proposed approach

The perceived speech quality $p_{\hat{s}}$ of the processed signal $\hat{s}(n)$ can be obtained from a listening test conducted with several assessors (see Section 4.2). This chapter focuses on two measures, proposed in [197] and [198], that aim at non-intrusively predicting $p_{\hat{s}}$, i.e., at computing an estimate $\hat{p}_{\hat{s}}$ of $p_{\hat{s}}$ from the signal $\hat{s}(n)$ while requiring neither a listening test nor a reference signal, i.e., $s(n)$ or $d_{\text{ref}}(n)$. This section first describes the used features, i.e., the extraction of MEs from the test signal before describing their combination with the considered predicting functions. In the following, "Tree" refers to the combination of MEs with a model tree and "LSTM" refers to the combination of MEs with the considered RNN.

4.1.1 *Considered features*

Both Tree and LSTM use modulation energies (MEs) as features, which have already been used in the field of speech quality prediction in [185] and have been further elaborated in [186]. The computation of these features is depicted in Fig. 4.1 and can be summarized as follows.

First, the signal under test $\hat{s}(n)$ is filtered by a gammatone filterbank with J channels, resulting in J filtered signals $\tilde{s}_j(n)$, where j denotes the filter index. The temporal envelope $e_j(n)$ is extracted from $\tilde{s}_j(n)$ as

$$e_j(n) = \sqrt{\tilde{s}_j^2(n) + \mathcal{H}\left\{\tilde{s}_j(n)\right\}^2}, \tag{4.1}$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. The temporal envelopes are divided into $L = \lceil (N-O)/(W-O) \rceil$ overlapping windowed frames using an overlap of O samples and a window of length W, with N the signal length. The modulation spectral energy $\overline{e}_j(k,\ell)$ is computed as the squared magnitude of the discrete Fourier transform of the ℓ-th frame in the k-th modulation frequency bin.

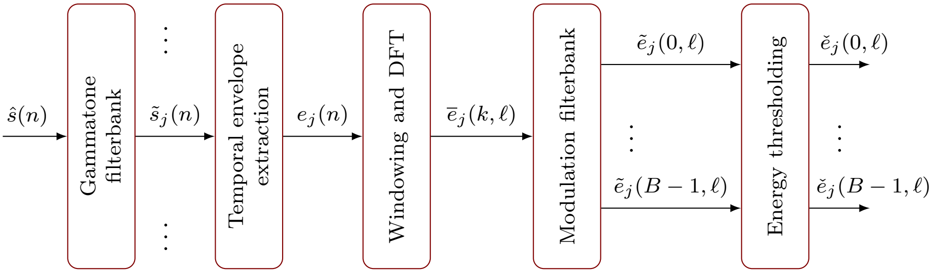

Fig. 4.1: Overview of the feature extraction. The test signal is filtered using a gammatone filterbank and the temporal envelope of each band is extracted. The modulation energies are extracted from these envelopes and energy thresholding is applied to compute the used features.

The modulation spectral energies $\overline{e}_j(k,\ell)$, for frequency bins in the interval k_{min} to k_{max}, are warped into B overlapping modulation bands whose centre frequencies are set as in [186], resulting in the warped modulation energies $\tilde{e}_j(b,\ell)$, where b denotes the index of the modulation band. As proposed in [186], thresholding is applied to $\tilde{e}_j(b,\ell)$, resulting in $\alpha e_{\text{peak}} \leq \check{e}_j(b,\ell) \leq e_{\text{peak}}$, where $0 \leq \alpha < 1$ and where

$$e_{\text{peak}} = \max_{j,b} \left(\frac{1}{L} \sum_{\ell=0}^{L-1} \tilde{e}_j(b,\ell) \right). \tag{4.2}$$

Finally, a feature vector $\mathbf{e}_\ell$ of length $J \cdot B$ is constructed for each ℓ-th time frame as

$$\mathbf{e}_\ell = \left[\check{e}_0(0,\ell), \ldots, \check{e}_0(B-1,\ell), \ldots, \check{e}_{J-1}(0,\ell), \ldots, \check{e}_{J-1}(B-1,\ell)\right]^{\mathrm{T}}. \tag{4.3}$$

An example of a sequence of MEs extracted from a short speech segment is depicted in Fig. 4.2.

In this thesis, as in [197,198], the different parameters of the feature extraction have been set as in [186]. The gammatone filterbank is applied to signals downsampled to 8 kHz and uses $J = 23$ channels with center frequencies ranging from 125 Hz to 4 kHz. The modulation frequency bins are grouped into $B = 8$ bands. The temporal envelope $e_j(n)$ is divided in frames using a Hamming window of length W corresponding to 256 ms and an overlap of length O corresponding to 224 ms. The indices k_{min} and k_{max} correspond to the range of modulation frequencies between 4 Hz and 40 Hz and α is set to lower bound modulation energies 30 dB below e_{peak}. These values have been shown to reduce the sensitivity of the extracted features to speakers and pitch content [186] compared to the settings initially proposed in [185].

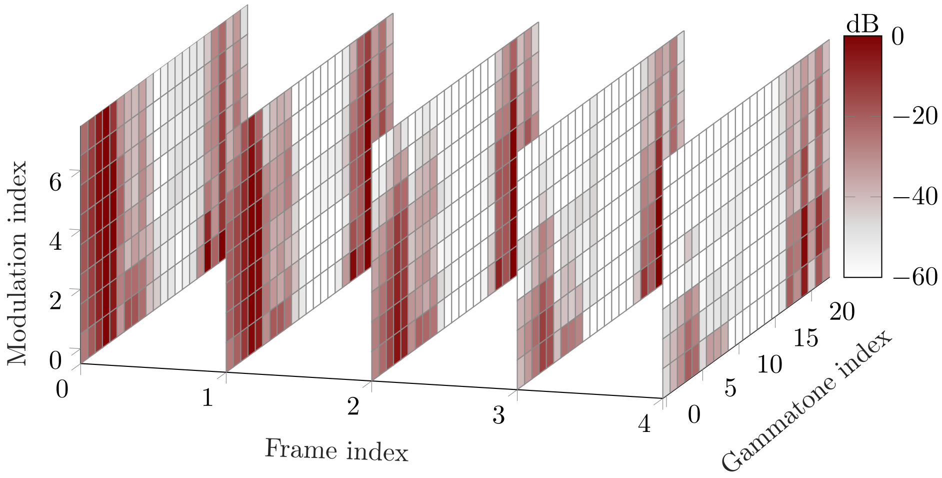

Fig. 4.2: Modulation energies extracted from 320 ms of clean speech. The coefficients in each frame are used to construct the sequence of vectors described by 4.3.

When using LSTM, the time ordered sequence described by (4.3) is used as input to the predicting function. However, when using Tree, as in previous use of the ME for speech quality prediction [185, 186], a single feature vector $\mathbf{e}$ of length B represents the signal $\hat{s}(n)$, i.e,

$$\mathbf{e} = \left[\mathsf{e}(0),\ \mathsf{e}(1),\ \ldots,\ \mathsf{e}(B-1)\right]^{\mathrm{T}}, \tag{4.4}$$

where

$$\mathsf{e}(b) = \frac{1}{JL}\sum_{j=0}^{J-1}\sum_{\ell=0}^{L-1}\check{e}_j(b,\ell). \tag{4.5}$$

The SRMR [185] and the $\mathrm{SRMR}_{\mathrm{norm}}$ [186] differ in the extraction of the vector $\mathbf{e}$ but both compute the estimate $\hat{p}_{\hat{s}}$ as the ratio between the lower and higher coefficients of $\mathbf{e}$. Tree computes the estimate $\hat{p}_{\hat{s}}$ as the output of a model tree, i.e., a combination of classification rules and regression but uses the same features as in [186] and therefore does not take into account time dependencies in the input signal.

4.1.2 *Model tree as predicting function*

The estimate $\hat{p}_{\hat{s}}$ of the perceived speech quality $p_{\hat{s}}$ is obtained by applying a predicting function $f(\cdot)$ to the feature vector $\mathbf{e}$, i.e. $\hat{p}_{\hat{s}} = f(\mathbf{e})$. In the case of both SRMR [185] and $\mathrm{SRMR}_{\mathrm{norm}}$ [186], this predicting function is equal to the energy ratio between the lower and upper modulation frequencies, i.e.

$$\mathrm{SRMR}_{\mathrm{norm}} = f_{\mathrm{SRMR}}(\mathbf{e}) = \frac{\sum_{b=0}^{b_{\mathrm{low}}-1}\mathsf{e}(b)}{\sum_{b=b_{\mathrm{low}}}^{b_{\mathrm{high}}-1}\mathsf{e}(b)}, \tag{4.6}$$

with b_{low} and and b_{high} denoting the lowest and highest considered modulation frequency index, respectively. As in [186], $b_{\mathrm{low}} = 4$ and b_{high} is determined as the index for which 90% of the modulation energy is accounted for, as proposed in [185]. When using Tree, the values of the vector $\mathbf{e}$ are normalized to range between 0 and 1. Though such normalization would have no influence on the output of (4.6) it is used for Tree as it is convenient to define decision rules.

Instead of using the energy ratio in (4.6) as the predicting function, Tree relies on a predicting function trained on a small corpus of (processed) speech data. It is hence assumed that a set of signals are available for training. In this thesis, we will use a model tree as the predicting function built using the so-called M5' algorithm [239], i.e., each leaf node of the decision tree consists of a linear model which predicts a value from the input features. The construction of the model tree requires a set of observations which are all associated with a feature vector and a target value. The set of observations here consists of the signals available in the training set, which are all associated with a feature vector $\mathbf{e}$ and a perceptual score $p_{\hat{s}}$. The model tree is built in two stages. In the first stage, a conventional binary decision tree is built by recursively splitting the set of observations into subsets in which the variation of the perceptual scores is maximal. This results in a large binary decision tree which can be used to predict a value for each observation. In the second stage, each node, starting from the top of the constructed decision tree, is replaced by a linear model if the application of this model results in a lower prediction error than the binary classification. The resulting model tree is finally simplified using pruning and smoothing as described in [239], and contains D binary decision nodes.

Although small values of D may result in underfitting, D should remain much smaller than the size of the training set to prevent overfitting, which would make the resulting predicting function inapplicable to signals not included in the training set. In addition, it should be noted that the predicting function $f(\cdot)$ resulting from the model tree is actually not designed to maximize the correlation in (2.13) but rather to minimize the mean squared error obtained on the training set. The effectiveness of the resulting predicting function, using the benchmark presented in Section 2.3, is discussed in Section 4.4.

4.1.3 *LSTM network as predicting function*

Artificial neural networks are usually composed of several layers. Each λ-th layer applies a non linear mapping to an input vector $\mathbf{x}^{\lambda}$, of length $L_{\mathbf{x}}^{\lambda}$, in order to compute an output vector $\mathbf{z}^{\lambda}$, of length $L_{\mathbf{z}}^{\lambda}$. It should be noted that, in this section, superscripts denote a layer index and not an exponent. This mapping is applied by multiplying a weight matrix $\mathbf{W}_{\mathbf{x},\mathbf{z}}^{\lambda}$ of size $L_{\mathbf{z}}^{\lambda} \times L_{\mathbf{x}}^{\lambda}$, where subscripts indicate the connections represented by the matrix, with the input vector $\mathbf{x}^{\lambda}$ before summing the results with a bias vector $\mathbf{b}_{\mathbf{z}}^{\lambda}$ of length $L_{\mathbf{z}}^{\lambda}$ and applying a non-linear activation function $\mathcal{F}(\cdot)$ to the result, i.e.,

$$\mathbf{z}^{\lambda} = \mathcal{F}\left(\mathbf{W}_{\mathbf{x},\mathbf{z}}^{\lambda}\mathbf{x}^{\lambda} + \mathbf{b}_{\mathbf{z}}^{\lambda}\right). \tag{4.7}$$

The values of $\mathbf{W}^{\lambda}_{\mathbf{x},\mathbf{z}}$ and $\mathbf{b}^{\lambda}_{\mathbf{z}}$ have to be learned during a training phase (see Section 4.3.2) and any number of layers can be used by setting $\mathbf{x}^{\lambda} = \mathbf{z}^{\lambda-1}$. Layers described by (4.7) and networks composed exclusively of such layers are commonly qualified as feed-forward networks.

The use of RNNs is a common extension of (4.7) to take time dependencies into account. Similarly as feed-forward artificial neural networks (ANNs), RNNs are composed of several layers. However, the input of the λ-th RNN layer is an ordered sequence $\mathbf{X}^{\lambda}$ of T^{λ} input vectors $\mathbf{x}^{\lambda}_{t}$, where $t \in [0, T^{\lambda} - 1]$ denotes the sequence index, i.e.,

$$\mathbf{X}^{\lambda} = \left\{\mathbf{x}^{\lambda}_{0}, \mathbf{x}^{\lambda}_{1}, \cdots, \mathbf{x}^{\lambda}_{T^{\lambda}-1}\right\}. \tag{4.8}$$

Each layer of an RNN computes a sequence $\mathbf{H}^{\lambda}$ of hidden vectors $\mathbf{h}^{\lambda}_{t}$ of length $L^{\lambda}_{\mathbf{h}}$ and a sequence $\mathbf{Z}^{\lambda}$ of output vectors $\mathbf{z}^{\lambda}_{t}$ of length $L^{\lambda}_{\mathbf{z}}$, both containing T^{λ} vectors and defined similarly as in (4.8). The vectors in these sequences are computed by iteratively applying [240]

$$\mathbf{h}^{\lambda}_{t} = \mathcal{F}\left(\mathbf{W}^{\lambda}_{\mathbf{x},\mathbf{h}}\mathbf{x}^{\lambda}_{t} + \mathbf{W}^{\lambda}_{\mathbf{h},\mathbf{h}}\mathbf{h}^{\lambda}_{t-1} + \mathbf{b}^{\lambda}_{\mathbf{h}}\right), \tag{4.9}$$

$$\mathbf{z}^{\lambda}_{t} = \mathcal{F}\left(\mathbf{W}^{\lambda}_{\mathbf{h},\mathbf{z}}\mathbf{h}^{\lambda}_{t} + \mathbf{b}^{\lambda}_{\mathbf{z}}\right), \tag{4.10}$$

where $\mathbf{W}^{\lambda}_{\mathbf{x},\mathbf{h}}$, $\mathbf{W}^{\lambda}_{\mathbf{h},\mathbf{h}}$ and $\mathbf{W}^{\lambda}_{\mathbf{h},\mathbf{z}}$ denote weight matrices of size $L^{\lambda}_{\mathbf{h}} \times L^{\lambda}_{\mathbf{x}}$, $L^{\lambda}_{\mathbf{h}} \times L^{\lambda}_{\mathbf{h}}$ and $L^{\lambda}_{\mathbf{z}} \times L^{\lambda}_{\mathbf{h}}$, respectively, and where $\mathbf{b}^{\lambda}_{\mathbf{h}}$ and $\mathbf{b}^{\lambda}_{\mathbf{z}}$ are bias vectors of length $L^{\lambda}_{\mathbf{h}}$ and $L^{\lambda}_{\mathbf{z}}$, respectively.

For our application, i.e. the prediction of speech quality, RNNs have two main advantages over feed-forward networks. First, the sequence of hidden vectors computed by an RNN allows the prediction to take into account temporal dependencies; second, the iterative updates can be applied to a sequence of arbitrary length. However, the values of the weight matrices and of the bias vectors still have to be learned during a training phase and the formulation in (4.9) and (4.10) can cause instability during training, leading to overly long training time or even divergence [241]. In order to avoid these issues, so-called gated units, such as in the LSTM layers used in this thesis, are used in practice.

Though in a standard RNN layer, the function $\mathcal{F}(\cdot)$ in (4.9) is commonly a simple non-linear function such as a sigmoid, in an LSTM layer, this function relies on iterative updates of sequences of vectors, $\mathbf{I}^{\lambda}$, $\mathbf{O}^{\lambda}$, $\mathbf{F}^{\lambda}$ and $\mathbf{C}^{\lambda}$, the so-called, input gate, output gate, forget gate and cell memory, respectively and their mutual influence on the layer's output is illustrated in Fig. 4.3. For each step t of the input sequence $\mathbf{X}^{\lambda}$, the vectors $\mathbf{i}^{\lambda}_{t}$ and $\mathbf{f}^{\lambda}_{t}$ are computed from the input vector $\mathbf{x}^{\lambda}_{t}$ and from the memory cell vector $\mathbf{c}^{\lambda}_{t-1}$ saved at the previous step, i.e.,

$$\mathbf{i}^{\lambda}_{t} = \mathcal{S}\left(\mathbf{W}^{\lambda}_{\mathbf{x},\mathbf{i}}\mathbf{x}^{\lambda}_{t} + \mathbf{W}^{\lambda}_{\mathbf{h},\mathbf{i}}\mathbf{h}^{\lambda}_{t-1} + \mathbf{W}^{\lambda}_{\mathbf{c},\mathbf{i}}\mathbf{c}^{\lambda}_{t-1} + \mathbf{b}^{\lambda}_{\mathbf{i}}\right), \tag{4.11}$$

$$\mathbf{f}^{\lambda}_{t} = \mathcal{S}\left(\mathbf{W}^{\lambda}_{\mathbf{x},\mathbf{f}}\mathbf{x}^{\lambda}_{t} + \mathbf{W}^{\lambda}_{\mathbf{h},\mathbf{f}}\mathbf{h}^{\lambda}_{t-1} + \mathbf{W}^{\lambda}_{\mathbf{c},\mathbf{f}}\mathbf{c}^{\lambda}_{t-1} + \mathbf{b}^{\lambda}_{\mathbf{f}}\right), \tag{4.12}$$

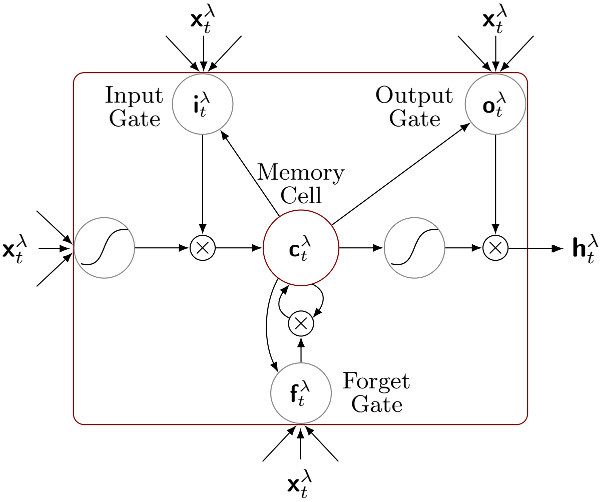

Fig. 4.3: Overview of the updates applied by an LSTM-layer. Updates are applied along the time-ordered sequence of input vectors resulting in a sequence of hidden vectors used to compute the output of the layer as in (4.10). The value of each hidden vector depends on the current input as well as on the memory cell and of the weights applied in the input, output and forget gates.

where $\mathcal{S}(\cdot)$ denotes the logistic sigmoid function. The resulting vectors $\mathbf{i}_t^\lambda$ and $\mathbf{f}_t^\lambda$ weight the influence of the current and previous input, respectively, to the updated vector $\mathbf{c}_t^\lambda$ computed as

$$\mathbf{c}_t^\lambda = \mathbf{f}_t^\lambda \mathbf{c}_{t-1}^\lambda + \mathbf{i}_t^\lambda \tanh\left(\mathbf{W}_{\mathbf{x},\mathbf{c}}^\lambda \mathbf{x}_t^\lambda + \mathbf{W}_{\mathbf{h},\mathbf{c}}^\lambda \mathbf{h}_{t-1}^\lambda + \mathbf{b}_{\mathbf{c}}^\lambda\right). \tag{4.13}$$

The influence of this memory cell vector $\mathbf{c}_t^\lambda$ to the layer output is weighted by the output gate $\mathbf{o}_t^\lambda$ computed as

$$\mathbf{o}_t^\lambda = \mathcal{S}\left(\mathbf{W}_{\mathbf{x},\mathbf{o}}^\lambda \mathbf{x}_t^\lambda + \mathbf{W}_{\mathbf{h},\mathbf{o}}^\lambda \mathbf{h}_{t-1}^\lambda + \mathbf{W}_{\mathbf{c},\mathbf{o}}^\lambda \mathbf{c}_t^\lambda + \mathbf{b}_{\mathbf{o}}^\lambda\right), \tag{4.14}$$

and used to compute the hidden vector $\mathbf{h}_t^\lambda$,

$$\mathbf{h}_t^\lambda = \mathbf{o}_t^\lambda \tanh\left(\mathbf{c}_t^\lambda\right), \tag{4.15}$$

from which the ouput vector $\mathbf{z}_t^\lambda$ is finally computed as per (4.10).

The LSTM measure uses the stacking of $\Lambda = 3$ layers as the predicting function of the quality of processed speech. Using this network structure, similar to the one used in [242] and depicted in Fig. 4.4, the speech quality of a signal $\hat{s}(n)$ is predicted by

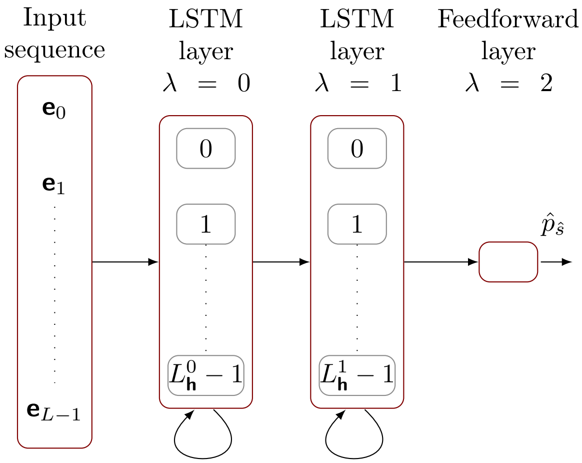

Fig. 4.4: Overview of the network used as predicting function. The first LSTM-layer computes a sequence of input vectors containing as many vectors as frames available in the test signal. The second LSTM-layer applies updates along this sequence and the last vector of its output sequence is input to the feed-forward layer whose sigmoid activation function results in a prediction bounded between 0 and 1.

using the sequence of L time ordered frames of features as input to the first LSTM layer, i.e. for $\lambda = 0$ we have $T^0 = L$, and

$$\mathbf{X}^0 = \left\{\mathbf{x}_0^0, \mathbf{x}_1^0, \cdots, \mathbf{x}_{T^0-1}^0\right\}, \text{ with,} \tag{4.16}$$

$$\mathbf{x}_t^0 = \mathbf{e}_t, \; t \in [0, T^0 - 1]. \tag{4.17}$$

The output sequence, obtained after iterating (4.11)–(4.15) and (4.10) along the input sequence, i.e, for all t, is used as input to a second LSTM layer, i.e., $\mathbf{X}^1 = \mathbf{Z}^0$ and the iterative updates of the second LSTM layer, $\lambda = 1$, yield the output sequence $\mathbf{Z}^1$. The last vector of this sequence is input to the last, feed-forward, layer, i.e., $\mathbf{x}^2 = \mathbf{z}_{T^1-1}^1$. Aiming at an estimate $\hat{p}_{\hat{s}}$ that is bounded between 0 and 1, we replace $\mathcal{F}(\cdot)$ in (4.7) by a sigmoid and compute the estimate of $p_{\hat{s}}$ as

$$\hat{p}_{\hat{s}} = \mathcal{S}\left(\mathbf{W}_{\mathbf{x},\mathbf{z}}^2 \mathbf{x}^2 + \mathbf{b}_{\mathbf{z}}^{\lambda}\right). \tag{4.18}$$

The values of the multiple weight matrices and bias vectors needed for the computation of (4.18) can be learned during a training phase using perceptually labeled training data. Section 4.2 presents the dataset that was collected for this purpose and the training procedure is described in Section 4.3.

4.2 Dataset

In order to train and evaluate the proposed measures, Tree and LSTM, described in Section 4.1, we collected a database of noisy and reverberant speech signals

Table 4.1: Expected behavior of the considered algorithms.

	UN	SC	MVDR	GWPE-MVDR
Denoising	Poor	Good	Fair	Fair
Dereverberation	Poor	Poor	Poor	Good
Speech distortions	Good	Fair	Good	Fair
Noise distortions	Good	Fair	Good	Poor

processed by several categories of algorithms and labeled in terms of perceived speech quality. This section provides short descriptions of the considered algorithms before describing the perceptual evaluation conducted in order to label signals in terms of perceived speech quality.

4.2.1 *Considered algorithms*

The algorithms considered in this chapter, as in the rest of the thesis, aim at computing the STFT $\hat{s}(k,\ell)$ of the enhanced speech signal $\hat{s}(n)$ from the M microphone channels $y_m(k,\ell)$. In addition to the unprocessed (UN) version of the signal, we considered three categories of algorithms, namely, single-channel spectral enhancement (SC) presented in Section 3.2 [195,196], the MVDR beamformer, as defined in Section 3.1, and the application of this beamformer to the output of the generalized weighted prediction error (GWPE) [136], denoted by GWPE-MVDR and introduced at the end of this subsection. These algorithms have been chosen for their applicability to realistic scenarios, e.g. real-time applications, as well as to provide a wide range of processing artefacts typically occurring in different reverberation and noise conditions. In the case of SC, the algorithm is similar to the single-channel described in Chapter 3 but is briefly presented here for completeness. The expected behavior of the algorithms in terms of interference reduction and introduced distortions is summarized in Table 4.1 and the categories of algorithms are briefly described in the next subsections.

All signals have a sampling frequency of f_s =16 kHz. Noisy and reverberant signals have been generated by convolving clean speech extracted from the WSJCAM0 database [202] with RIRs and adding noise to the resulting reverberant speech. We used 3 different RIRs extracted from the ACE challenge dataset [79] recorded using a 42 cm linear array of $M = 8$ equidistant microphones, whose characteristics, summarized in Table 4.2, have been selected to represent a wide range of reverberation levels. We considered two noise types, namely *fan* noise and *babble*, for which noise signals recorded in the same rooms and with the same microphones positions as for the RIRs are available. We consider two SNRs, namely of 5 dB and 15 dB, calculated according to [243]. The 12 resulting combinations of RIRs with noise types and SNRs will be referred to as *acoustic conditions* in the remainder of this

Table 4.2: RIRs used to generate the perceptually evaluated signals along with their respective characteristics.

Labels	Room	T_{60} [s]	DRR [dB]
RIR 1	Office 1	0.35	10.45
RIR 2	Building Lobby	0.77	5.13
RIR 3	Lecture Room 2	1.26	3.79

chapter. When referring to UN as an algorithm, we consider $\hat{s}(n) = y_{\text{ref}}(n)$ with ref arbitrarily set to 1.

4.2.1.1 *Single-channel spectral enhancement (SC)*

SC algorithms estimate $\hat{s}(k, \ell)$ by applying a real valued gain to the STFT of one of the input channel (see Fig. 3.2), i.e.,

$$\hat{s}(k, \ell) = g(k, \ell) y_{\text{ref}}(k, \ell). \tag{4.19}$$

The gain $g(k, \ell)$ is computed as

$$g(k, \ell) = \max\left(\tilde{g}(k, \ell), g_{\min}\right), \tag{4.20}$$

where $g_{\min}$ is, as in Subsection 3.2.1, a spectral floor introduced to limit possible speech distortion and where $\tilde{g}(k, \ell)$ is in this chapter computed as the solution to the MMSE estimator of the speech amplitude proposed in [215] and described from (3.22) to (3.24). As described in Section 3.2, the computation of the spectral gains relies on various PSDs that, in practice, are unknown and their estimates, $\hat{\sigma}^2_{d\text{ref}}(k, \ell)$, $\hat{\sigma}^2_{r\text{ref}}(k, \ell)$ and $\hat{\sigma}^2_{v\text{ref}}(k, \ell)$ have to be used instead.

The choice of the PSD estimators can greatly influence the performance of SC algorithms. In this chapter, we denote by "SCa" the combination described in Chapter 3 [196], which has been shown effective in improving speech quality in reverberant scenarios [96, 162]. The estimates of the PSDs, $\sigma^2_{v\text{ref}}(k, \ell)$, $\sigma^2_{r\text{ref}}(k, \ell)$ and $\sigma^2_{d\text{ref}}(k, \ell)$, are estimated using a modified version of the well-known MS estimator [90], the Lebart approach [87] and cepstral smoothing [216], respectively, $\tilde{g}(k, \ell)$ is computed using (3.22) and $g_{\min}$ is set to a minimum gain of -10 dB. A detailed description of the approach is available in [196].

Aiming at measuring the effect of distortions that a poorly tuned SC algorithm could introduce, we used a modified version of this scheme denoted by "SCb". In this case, we estimate $\sigma^2_{v\text{ref}}(k, \ell)$ and $\sigma^2_{d\text{ref}}(k, \ell)$ using the estimators proposed in [94] and in [98], respectively, and set $g_{\min}$ to a minimum gain of -30 dB.

4.2.1.2 *MVDR beamformer*

The MVDR beamformer considered in this chapter estimates $\hat{s}(k,\ell)$, as presented in Section 3.1, by filtering and summing the STFT coefficients of the multichannel input

$$\hat{s}(k,\ell) = \mathbf{w}_{\hat{\theta}}^{\mathrm{H}}(k)\mathbf{y}(k,\ell) \tag{4.21}$$

where $\mathbf{w}_{\hat{\theta}}(k)$ denotes the stacked filter coefficient vector of the beamformer, see (3.7), and where $\mathbf{y}(k,\ell)$ denotes the M-dimensional stacked vector of the received microphone signals, see (3.2). The computation of $\mathbf{w}_{\hat{\theta}}(k)$ requires an estimate $\hat{\mathbf{\Gamma}}(k)$ of the noise covariance matrix and a steering vector $\mathbf{d}_{\hat{\theta}}(k)$ steered towards the estimated DOA $\hat{\theta}$.

In this chapter, the estimate $\hat{\mathbf{\Gamma}}(k)$ is computed as

$$\hat{\mathbf{\Gamma}}(k) = \overline{\mathbf{\Gamma}}(k) + \varrho(k)\mathbf{I}_M, \tag{4.22}$$

where $\overline{\mathbf{\Gamma}}(k)$ denotes the coherence matrix of a diffuse noise field [213], $\mathbf{I}_M$ denotes the $M \times M$-dimensional identity matrix and $\varrho(k)$ denotes a frequency-dependent regularization parameter used to limit potential amplification of uncorrelated noise, especially at low frequencies. This regularization parameter is computed iteratively such that

$$\mathbf{w}_{\hat{\theta}}^{\mathrm{H}}(k)\mathbf{w}_{\hat{\theta}}(k) \leq \mathrm{WNG}_{\max}, \tag{4.23}$$

where $\mathrm{WNG}_{\max}$ denotes the so-called white noise gain constraint [218]. In this chapter we set this constraint to -10 dB, compute the steering vector $\mathbf{d}_{\hat{\theta}}(k)$ using a far-field assumption and measure the true θ from the main peaks of the used RIRs. In order to evaluate the impact of steering error on the performance of the beamformer we consider perfectly steered, i.e. $\hat{\theta} = \theta$, denoted by "MVDRa", and missteered beamformer, i.e. $\hat{\theta} = \theta + \epsilon_{\hat{\theta}}$ with $\epsilon_{\hat{\theta}} = \pi/4$ denoted by "MVDRb".

4.2.1.3 *Combination of GWPE and MVDR*

The combination of GWPE and MVDR beamformer (GWPE-MVDR) considered in this thesis estimates $\hat{s}(k,\ell)$ as

$$\hat{s}(k,\ell) = \mathbf{w}_{\hat{\theta}}^{\mathrm{H}}(k)\left(\mathbf{y}(k,\ell) - \hat{\mathbf{r}}(k,\ell)\right), \tag{4.24}$$

where $\mathbf{w}_{\hat{\theta}}(k)$ is computed as in (3.7) and

$$\hat{\mathbf{r}}(k,\ell) = [\hat{r}_1(k,\ell),\ \hat{r}_2(k,\ell),\ \dots\ ,\hat{r}_M(k,\ell)]^{\mathrm{T}}, \tag{4.25}$$

where $\hat{r}_m(k,\ell)$ denotes an estimate of $r_m(k,\ell)$, i.e., $\hat{s}(k,\ell)$ is estimated by subtracting a complex valued estimate of the late reverberation from the multichannel input signal before applying an MVDR beamformer.

In this thesis, the estimate $\hat{\mathbf{r}}(k,\ell)$ is computed using the approach described in [136], i.e. as

$$\hat{\mathbf{r}}(k,\ell) = \mathbf{P}^{\mathrm{H}}(k,\ell)\tilde{\mathbf{y}}(k,\ell-\Delta), \tag{4.26}$$

where Δ denotes a delay introduced to preserve the early reflections, and

$$\mathbf{P}(k,\ell) = [\mathbf{p}_1(k,\ell),\ \cdots,\ \mathbf{p}_M(k,\ell)] \in \mathbb{C}^{ML_P \times M}, \tag{4.27}$$

where $\mathbf{p}_m(k,\ell) \in \mathbb{C}^{ML_P}$ denotes the vector of the prediction filter coefficients applied to the m-th channel, and where

$$\begin{aligned}\tilde{\mathbf{y}}(k,\ell) =& [y_1(k,\ell),\ \cdots,\ y_1(k,\ell - L_P + 1)\ ,\cdots,\\ & y_M(k,\ell),\ \cdots\ ,y_M(k,\ell - L_P + 1)]^{\mathrm{T}},\end{aligned} \tag{4.28}$$

denotes a vector of STFT coefficients of length $M \cdot L_P$.

For each time-frequency bin, the matrix $\mathbf{P}(k,\ell)$ is computed by applying γ iterative updates aiming at solving the optimization problem [136]

$$\begin{aligned}&\underset{\mathbf{P}(k,\ell)}{\operatorname{argmin}}\ \operatorname{tr}\ \left\{\mathbf{P}^{\mathrm{H}}(k,\ell)\hat{\mathbf{A}}(k,\ell)\mathbf{P}(k,\ell)\right\} - 2\Re\left\{\operatorname{tr}\ \left\{\mathbf{P}^{\mathrm{H}}(k,\ell)\hat{\mathbf{B}}(k,\ell)\right\}\right\}\\ &\text{subject to } |\mathbf{P}^{\mathrm{H}}(k,\ell)\tilde{\mathbf{y}}(k,\ell-\Delta)|^2 \leq \hat{\boldsymbol{\sigma}}_r^2\,(k,\ell)\,,\end{aligned} \tag{4.29}$$

where

$$\hat{\boldsymbol{\sigma}}_r^2\,(k,\ell) = [\hat{\sigma}_{r1}^2\,(k,\ell)\,,\ \cdots,\ \hat{\sigma}_{rM}^2\,(k,\ell)]^{\mathrm{T}}, \tag{4.30}$$

and $\hat{\sigma}_{rm}^2\,(k,\ell)$ is computed similarly as in Section 4.2.1.1 and with

$$\hat{\mathbf{A}}(k,\ell) = \sum_{i=1}^{\ell} \delta^{\ell-i}\hat{w}_P(k,i)\tilde{\mathbf{y}}(k,i-\Delta)\tilde{\mathbf{y}}^{\mathrm{H}}(k,i-\Delta), \tag{4.31}$$

$$\hat{\mathbf{B}}(k,\ell) = \sum_{i=1}^{\ell} \delta^{\ell-i}\hat{w}_P(k,i)\tilde{\mathbf{y}}(k,i-\Delta)\mathbf{y}^{\mathrm{H}}(k,i), \tag{4.32}$$

where $\delta \in [0,1]$ denotes a smoothing constant, and $\hat{w}_P(k,\ell)$ denotes the weight used to emphasize frames where the the signal to be preserved is expected to have low power, computed as

$$\hat{w}_P(k,\ell) = \left(\frac{1}{M}\|\hat{\boldsymbol{\sigma}}_d^2\,(k,\ell)\,\|_2^2 + \epsilon\right)^{-1}, \tag{4.33}$$

where ϵ denotes a small regularization constant and where

$$\hat{\boldsymbol{\sigma}}_d^2\,(k,\ell) = [\hat{\sigma}_{d1}^2\,(k,\ell)\,,\ \cdots,\ \hat{\sigma}_{dM}^2\,(k,\ell)]^{\mathrm{T}}, \tag{4.34}$$

where $\hat{\sigma}_{dm}^2\,(k,\ell)$ is an estimate of $\sigma_{dm}^2(k,\ell)$ computed from $\hat{\sigma}_{rm}^2\,(k,\ell)$ and $\hat{\sigma}_{ym}^2\,(k,\ell)$ using recursive temporal smoothing.

It can be noted that the optimization problem in (4.29) does not take noise into account as the approach presented in [136] has been designed aiming at dereverberation in noise-free scenarios. The filtered noise signal resulting from (4.24) might have different spatial properties than the noise signal recorded by the microphones and might result in lower noise reduction achieved by GWPE-MVDR compared to

Table 4.3: Overview of the dataset of perceptually evaluated signals, excluding references and anchors. Numbers denote the duration in minutes, for each combination of acoustic condition and algorithm, for a total of 5.23 hours and 1920 perceptually evaluated signals. Properties of the considered RIRs are summarized in Table 4.2.

			UN	SC		MVDR		GWPE-MVDR	
	Noise	SNR		a	b	a	b	a	b
RIR 1	Fan	5 dB	6.27	3.55	3.03	3.41	3.31	3.35	3.45
		15 dB	6.32	3.04	3.14	3.41	3.49	3.14	3.20
	Babble	5 dB	6.52	3.38	3.13	3.27	3.30	3.14	3.67
		15 dB	6.36	3.50	2.96	3.00	3.44	3.04	3.20
RIR 2	Fan	5 dB	6.53	3.22	2.92	3.06	3.40	2.87	3.17
		15 dB	6.63	2.94	3.19	3.05	3.45	3.21	3.52
	Babble	5 dB	6.52	3.43	3.37	3.30	3.51	3.35	3.26
		15 dB	6.62	3.58	3.27	3.30	3.30	3.25	3.32
RIR 3	Fan	5 dB	6.88	3.40	3.26	3.34	3.33	3.55	3.31
		15 dB	6.64	3.29	3.27	3.09	3.44	3.06	3.12
	Babble	5 dB	6.44	3.41	3.24	3.18	3.04	3.12	3.31
		15 dB	6.48	3.28	3.46	3.20	3.32	3.50	3.51

MVDR alone. In practice, GWPE-MVDR could be combined with spectral suppression to overcome such drawbacks. Such combination has not been considered in this chapter in order to obtain processed signals containing a wide range of processing artefacts. We used prediction filters of length $L_P = 5$ and a smoothing constant $\delta = 0.95$. Other parameters have been set as in [136]. Similarly as for MVDR, we consider both perfect steering and steering error and refer to the corresponding settings as "GWPE-MVDRa" and "GWPE-MVDRb", respectively.

All STFTs have been computed using a Hamming window. In the case of SC and MVDR, we used a window of 32 ms and an overlap of 16 ms while in the case of GWPE-MVDR, we used a window of 64 ms and an overlap of 48 ms, in order to replicate the implementation from [196] and [136].

4.2.2 *Perceptual evaluation*

In order to obtain a dataset of processed signals labeled in terms of *overall quality*, we conducted a MUSHRA test [37] involving 20 self-reported normal-hearing assessors.

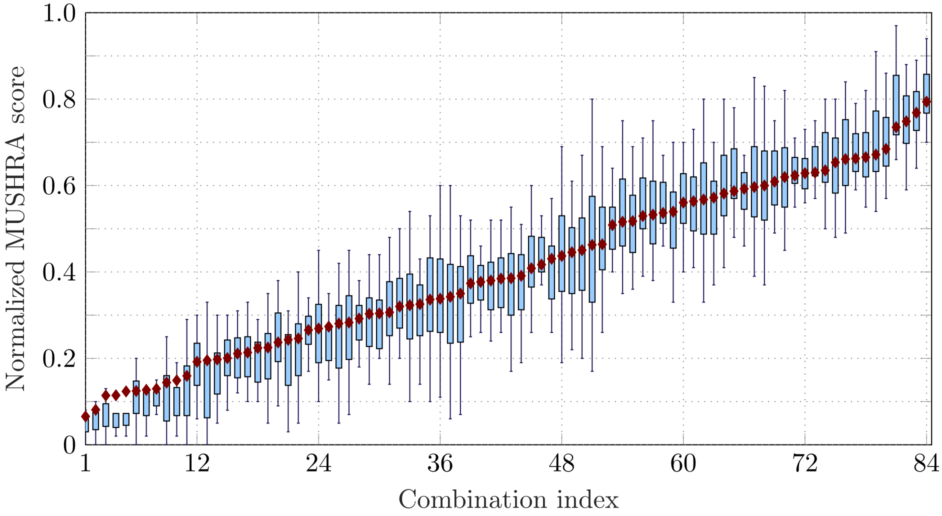

Fig. 4.5: MUSHRA scores, normalized to range between 0 and 1, for all combinations of acoustic conditions, algorithms and settings, excluding references and anchors. The mean, represented by a diamond, is considered the ground truth.

All assessors evaluated all combinations of the algorithm categories SC, MVDR and GWPE-MVDR, with two settings being considered for each, e.g., SCa and SCb, with the 12 acoustic conditions described in Section 4.2.1.

For each assessor, this total of 72 combinations was divided into two equally sized groups assigned to two sessions of listening tests. For each session, the UN algorithm was added to the group of combinations to be evaluated resulting in a total of 48 combinations per session and per assessor. The 48 combinations were randomly divided into six partitions and three clean male speech and three clean female speech utterances were randomly extracted from the WSJCAM0 database [202], with a sanity check insuring that no utterance would have been previously assigned to another session or assessor. One clean speech utterance was randomly assigned to each partition and used to generate signals, for the corresponding combinations as described in Section 4.2.1. Each partition was appended with the clean signal, used as reference signal, and two anchors, differing in the type of considered noise signal, either *babble* or *fan* noise. These anchors were generated by convolving the clean signal with the first channel of RIR 3 (see Table 4.1), and adding noise with a SNR of 5 dB measured according to [243] and band-pass filtering the resulting noisy and reverberant signal according to [244]. Once all signals were generated, the test procedure for each assessor was conducted as follows.

The signals under test were normalized to have their maximum level, calculated over segments of 500 ms, equal to 65 dB SPL. Stimuli were diotically presented over headphones (Senheiser HD200) in a soundproof booth. The speech material presented to the assessor was first presented in a training phase during which the assessor could listen to all stimuli in order to become familiar with the material.

Subsequently, the signals corresponding to each partition of test conditions were presented simultaneously on a screen. For each signal, there was a corresponding slider that the assessor was prompted to use in order to grade the overall quality of the material on an integer-valued scale between 0 (poor) and 100 (excellent). It can be noted that these scores are divided by 100 to obtain the score, bounded between 0 and 1, that the methods presented in Section 4.1 aims at predicting. The scores assigned to the reference and anchor conditions were used to ensure that the assessors conducted the task reliably, i.e., that they assigned the highest score to the hidden reference and a low score to the anchor. The significance of differences between the 12 acoustic conditions was assessed using a repeated-measures analysis of variance (ANOVA) and post-hoc analysis, whose details are presented in Appendix A. An overview of the collected dataset is presented in Table 4.3 and the scores for all combinations are summarized in Fig. 4.5. In the remainder of this chapter, we consider the true perceived speech quality of a signal to be the mean of the scores assigned by the assessors to all signals of the same acoustic condition, algorithm and algorithm setting and refer to it as the *ground truth.*

4.3 Experiments

4.3.1 *Benchmark and figures of merit*

The results presented in Section 4.4 compare the performance of the proposed measures, e.g., Tree and LSTM, with several measures taken from the literature and included already in the benchmark described in Chapter 2. Though aiming at non-intrusive prediction of the speech quality, this benchmark includes three intrusive measures, namely PESQ [176], POLQA [177] and PEMO-Q [178] as they are commonly used to evaluate speech enhancement algorithms. It should however be emphasized that, as the computation of these measures requires the clean reference signal, they are not applicable in all scenarios and have an advantage in terms of performance compared to non-intrusive measures. Our benchmark includes three non-intrusive measures, namely P.563 [184], ANIQUE+ [192] and $\text{SRMR}_{\text{norm}}$ [186]. It should be noted that both Tree and LSTM rely on predicting functions trained using machine-learning techniques and that, contrary to the other considered measures, their performance depends on the data included in the training set.

We assess the performance of the proposed measure and of the benchmark measures using the four figures of merit described in Section 2.3. For each measure, the linear relationship between the predicted quality and the ground truth is quantified using the Pearson correlation coefficient ρ, the ranking capability of each measure is quantified by the Spearman rank correlation coefficient ρ_{spear} and the correlation coefficient ρ_{sig} is computed similarly as ρ after applying a sigmoidal mapping, whose parameters are computed from the training set, to the predicted values as described in Section 2.3. Finally, ϵ-RMSE is used to represent the error between the predicted value and the ground truth. This figure of merit is similar to the conventional RMSE but takes the uncertainty of the subjective ratings into account, i.e., ϵ-RMSE will be lower if the variance of the subjective ratings is high. An ideal measure should yield correlation values close to one and an ϵ-RMSE close to zero. Details on the computation of ρ_{sig} and ϵ-RMSE can be found in [40].

4.3.2 *Training framework*

The training of the predicting functions used by both Tree and LSTM, as well as the linear mapping used in the computation of ρ_{sig} require a training set of signals for which the ground truth value of $p\hat{s}$ is known. In addition, a test set is needed to assess the performance of these trained measures and of the benchmark measures listed above. The network described in Section 4.1 was set with $L_{\mathbf{h}}^0 = L_{\mathbf{h}}^1 = 128$ and was trained using Keras [245] and the Adam algorithm [246]. Influence of the network size is examined in Appendix B. During training, zero padding was applied to ensure that all sequences had same length and a masking layer was added before the first LSTM layer to ignore time frames containing only zeros. Each training epoch computed as many iterations as needed to take the entire training set into account using a batch size of 128 sequences. In our implementation dropout [247] was applied both to the input of the network and to the output of each LSTM layer, i.e., 30% of the values input to the network and output by each LSTM layer were randomly selected and replaced by zeros at each iteration. At each iteration, the model, i.e. weight matrices and bias vectors, was updated to minimize the mean squared error (MSE) between the predicted and ground truth value of the speech quality assigned to each file of the training set. In order to avoid overfitting, 10% of the training set was set aside prior to each training phase to be used as a validation set. The training algorithm computed 500 epochs and testing was done using the model that yielded the lowest MSE over the validation set.

We conducted three experiments that differ in the training and testing sets constructed from the dataset presented in Section 4.2. In all experiments, anchors and reference signals were discarded before training and testing. The first experiment aims at assessing the ability of the proposed measures to predict the speech quality from signals processed using a single category of algorithms, e.g., SC, but different settings, e.g. SCa and SCb. For this purpose, the dataset was divided into 4 subsets containing only files processed with the same category of algorithm (UN, SC, MVDR and GWPE-MVDR). For each subset, we used 5-fold cross validation, proceeding as follows. The 20 assessors have been randomly divided into 5 equally-sized disjoint groups. For each fold, the signals in one of these groups were considered as the test set, while the data corresponding to the remaining groups were used for training. Using this folding, assessors, speech stimuli and noise segments always differ between training and testing. The second experiment aims at assessing the ability of the proposed measures to predict the speech quality from signals processed with various categories of algorithms. For this purpose, we use the same 5-fold cross-validation procedure but apply it to the entire collected dataset. We used the same training sets and folds for LSTM, Tree, and learning of the parameters of the sigmoid used in the computation of ρ_{sig}.

The third experiment examines the behavior of the proposed measures in case of mismatch between the algorithms included in the training and the testing set. For

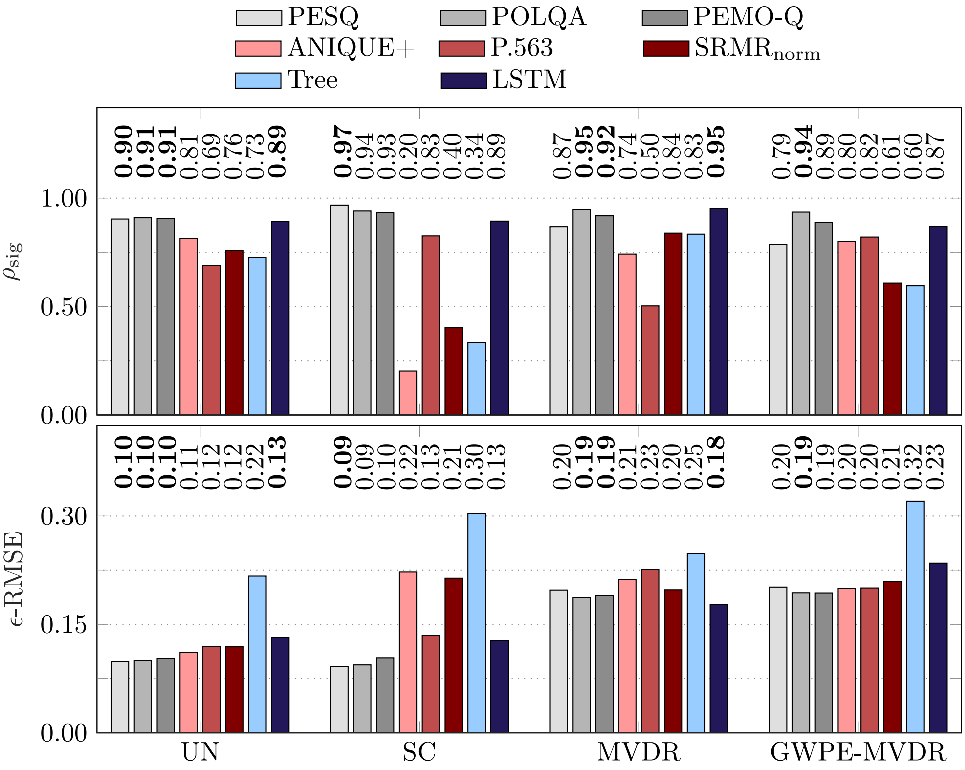

Fig. 4.6: Performance of the considered measures in terms of ρ_{sig} (top) and ϵ-RMSE (bottom). The labels along the x-axis denote the category of algorithms included in training and testing sets. Numbers in bold typeface denote the best attained performance (statistically indifferent) per considered category of algorithm.

this purpose, all signals processed with a single category of algorithms are included in the testing set while all others are included in the training set. It should be noted that using such partition yields a larger training set than for the previous experiments.

The figures of merit reported for the first and second experiments in Section 4.4 are averaged over all five folds. In the case of the correlation measures, a Fisher Z-test, at a significance level of 0.05, has been conducted before averaging to ensure that the values did not differ significantly between folds [248]. A similar Fisher Z-test has been used to determine if the difference between the correlations measures yielded by the considered measures were significant. In the case of ϵ-RMSE, significance was determined using the F-measure criteria suggested by ITU-T in [249] and detailed, e.g., in [40].

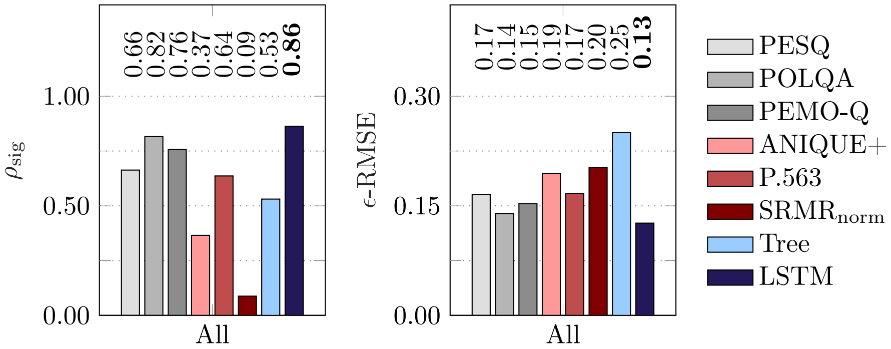

Fig. 4.7: Performance of the considered measures in terms of ρ_{sig} (left) and ϵ-RMSE (right). All considered categories of algorithms were included in both training and testing sets. Numbers in bold typeface denote the best attained performance (statistically indifferent).

4.4 Results

This section reports the results obtained considering the different training and testing sets previously described. As the three measures of correlation showed similar behavior at all of the considered measures we only report ρ_{sig} and ϵ-RMSE.

The performance obtained when training and testing the proposed measure for a single category of algorithms at a time are depicted in Fig. 4.6 along with the performance of the considered benchmark measures on the same testing sets. With the exception of LSTM, non-intrusive measures are consistently outperformed by intrusive measures, as could be expected. LSTM, however, yields similar performance as the intrusive measures for all considered categories of algorithms and, when training and testing on either unprocessed signals (UN) or signals processed using the MVDR beamformer, there is no significant difference between LSTM and the intrusive measures (indicated by bold typeface in Fig. 4.6). Although for SC and GWPE-MVDR, LSTM yields a slightly poorer performance than the intrusive measures, it outperforms all non-intrusive measures, including Tree, in terms of both ρ_{sig} and ϵ-RMSE, except for GWPE-MVDR where LSTM yields a slightly higher ϵ-RMSE than the benchmark measures.

The non-intrusive benchmark measures yield similar performance for unprocessed signals but perform inconsistently across the other categories of algorithms. Notably, ANIQUE+, SRMR$_{\text{norm}}$ and Tree yield low ρ_{sig} (0.2 to 0.4) and high ϵ-RMSE (0.2 to 0.3) in the case of SC. As both SRMR$_{\text{norm}}$ and Tree use ME features which are, contrary to the case of LSTM, averaged over time, this suggests that taking into account the time-dependency is beneficial. It can be noted than the difference in performance along different categories of algorithms is coherent with the results from previous works such as [188], in which it appeared that existing quality measures are often reliable when considering only one category of algorithms.

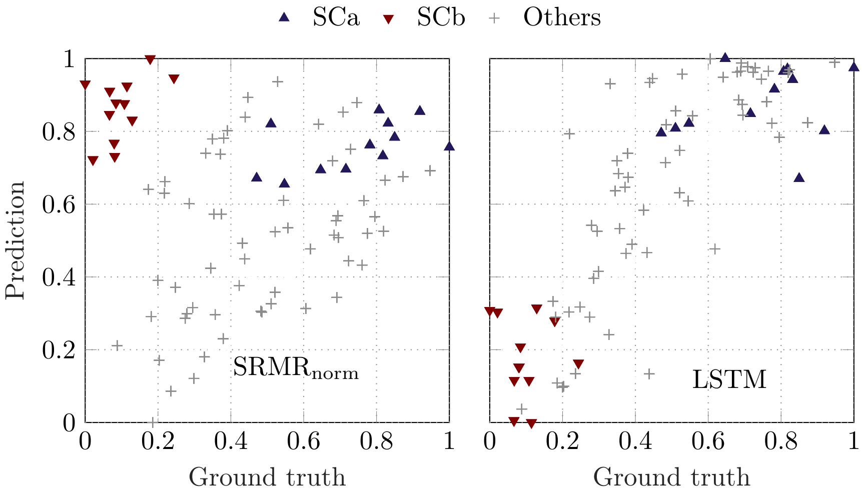

Fig. 4.8: Scatter plot of the predicted speech quality over ground truth data for SRMR$_{\text{norm}}$ (left) and the LSTM-based measure (right) when trained and tested for all considered algorithm categories, for all signals. Values corresponding to signals processed using SCa and SCb are highlighted, for readability, scores are normalized to range between 0 and 1.

The performance obtained when training and testing Tree and LSTM for all categories of algorithms are depicted in Fig. 4.7 along with the performance of the considered benchmark measures on the same testing sets. Unsurprisingly, for all measures, correlations are lower than when considering a single category at a time and, still with the exception of the proposed measure, non-intrusive measures are consistently outperformed by intrusive measures. The measure Tree performs poorly suggesting that, though it had been shown to yield high correlations in [197], it is not suitable to predict the speech quality in various acoustic conditions or with algorithms that might produce high levels of distortions. Similarly as in the case of SC, shown in Fig. 4.6, SRMR$_{\text{norm}}$ performs poorly with ρ_{sig} =0.09, while the proposed LSTM-based measure outperforms all measures, including the intrusive ones, in terms of both ρ_{sig} and ϵ-RMSE. This behavior is better illustrated in Fig. 4.8. In terms of ground truth, a clear divide appears between SCa and SCb, illustrating the clear preference of assessors for the well tuned single-channel scheme (SCa) over the purposefully poorly tuned (SCb). It appears as well that SRMR$_{\text{norm}}$ largely overestimates the speech quality for signals processed with SCb while the proposed measure is able to adequately reflect the difference in performance between the two settings. This behavior is to be expected as, by averaging features over time, SRMR$_{\text{norm}}$ effectively discards information about time-varying distortions (such as musical noise) while the proposed measure is designed to model such time-dependent effects.

As both Tree and the proposed LSTM-based measure are dependent on a training phase, one might want to consider the performance obtained in case of mismatch between the algorithms included in the training and the testing set. The performance

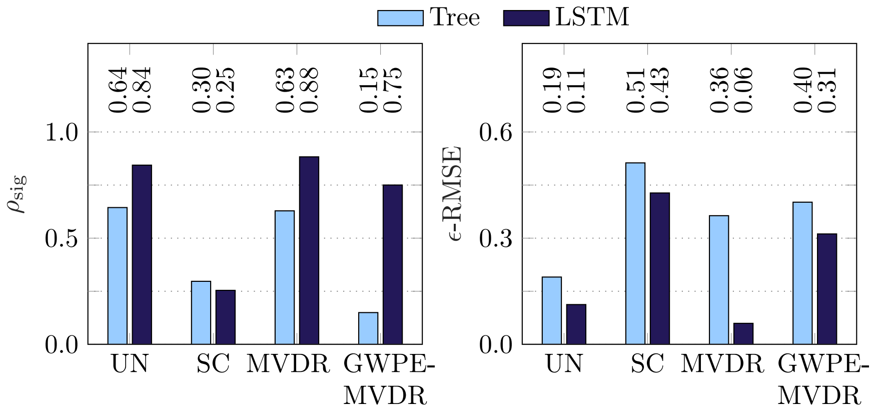

Fig. 4.9: Performance of the considered measures in terms of ρ_{sig} (left) and ϵ-RMSE (right). The labels along the x-axis denote the category of algorithms included in testing sets while all other categories were included in training.

of Tree and of LSTM in the presence of such mismatch is depicted in Fig. 4.9. It appears that when using a testing set composed of signals processed using either UN or MVDR, both Tree and LSTM yield similar performance as in the previous experiments in terms of ρ_{sig} and an even lower ϵ-RMSE. Such behavior can be explained by the fact that MVDR, even misteered, does not introduce large amount of distortions and that the predicting functions were trained on a larger training set. However, mismatch between algorithms included in training and testing greatly deteriorates performance of both Tree and LSTM when the testing set is composed of signals processed using either GWPE-MVDR or SC.

In the case of GWPE-MVDR, for LSTM, correlation decreases only slightly in comparison with previous experiments, $\rho_{\text{sig}} = 0.75$. However, the variance is high, ϵ-RMSE$= 0.31$. In the case of SC, both Tree and LSTM fail in their prediction, with low correlations $\rho_{\text{sig}} = 0.30$ and $\rho_{\text{sig}} = 0.25$ for Tree and LSTM, respectively. This difference in performance between the two measures and algorithms considered for testing is better illustrated in Fig. 4.10.

It appears that in the case of GWPE-MVDR, Tree does not yield an accurate prediction for any of the settings, i.e., GWPE-MVDRa and GWPE-MVDRb, while LSTM seems to overestimate the quality of signals processed with GWPE-MVDRa. In the case of SC, Tree fails similarly as in the GWPE-MVDR case but the poor performance of LSTM seems to come from an overestimation of the quality of signals processed using SCb. This behavior is unsurprising considering that the distortions introduced by spectral enhancement, e.g., musical noise, differ greatly from the ones introduced by the other algorithms and that, using this mismatch training set, the predicting functions could not take them into account. Consequently, neither Tree nor LSTM can be reliably used if the algorithm category under test is not included in the training set.

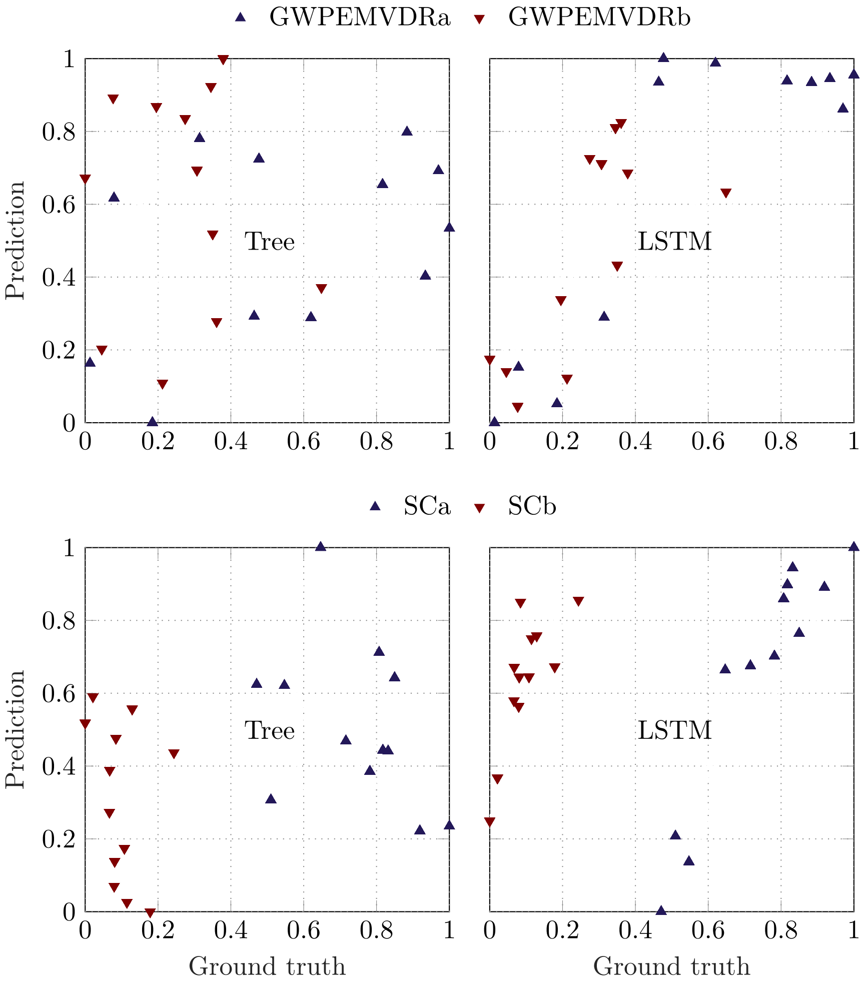

Fig. 4.10: Scatter plot of the predicted speech quality over ground truth data for Tree (left) and the LSTM-based measure (right) when using a testing set of signals processed using either GWPE-MVDR (top) or SC (bottom). Scores are normalized to range between 0 and 1.

4.5 Summary

Aiming at non-intrusively predicting the quality of processed signals, in this chapter, we examined the combination of modulation energies with a model tree and with an RNN with LSTM cells that takes the time-dependency of the target signal into account. For this purpose, we collected a large dataset of perceptually evaluated signals representing a wide range of acoustic conditions and various categories of algorithms with different settings. We conducted several experiments, differing in terms of training and testing sets used to train and evaluate the proposed measures. The aim of these experiments was to evaluate the reliability of the proposed measures when trained and tested for either a single category of algorithms or several categories, and to investigate the performance of the measures in case of a significant mismatch between the algorithms included in the training and the testing sets.

Experimental results show that LSTM is the most promising measure and that, when trained and tested for a single category of algorithms, LSTM outperforms the considered non-intrusive benchmark measures and yields a similar performance as the intrusive benchmark measures. When trained and tested for several categories of algorithms, LSTM even outperforms both intrusive and non-intrusive benchmark measures. However, as could be expected, both Tree and LSTM can be unreliable in case of a mismatch in terms of algorithms between the training and testing sets. Consequently, neither Tree nor LSTM are suitable to assess the performance of completely new categories of algorithms. However, LSTM could be a useful approach for the real-time selection of algorithms or algorithm parameters.

5

CONCLUSION AND FURTHER RESEARCH

This chapter provides a summary of the main contributions of this thesis and presents possible directions that could be explored to extend this work.

5.1 Conclusion

Many speech-based applications rely on recording the speech of a distant user using one or more microphones. Such applications comprise, e.g., hands-free communication systems, voice-controlled home assistants or hearing aids. In most cases, the recorded signals are corrupted by both noise, caused by undesired sound sources as well as reverberation, caused by reflections of the target speech. Both noise and reverberation can severely degrade the performance of speech-based applications, e.g., reduce the speech recognition accuracy of ASR systems or degrade the speech quality in speech communication systems.

This thesis addresses two important aspects of speech-based applications. First, the enhancement of speech signals recorded in the presence of both noise and reverberation in order to improve the robustness of ASR systems and improve the speech quality in speech communication systems. Second, the evaluation of speech enhancement algorithms that aim at reducing noise and reverberation while improving the speech quality.

In Chapter 3, we proposed a speech enhancement algorithm for joint noise and reverberation reduction. This algorithm can be applied in single-channel scenarios and in multichannel scenarios by including an MVDR beamformer. The main contribution of this chapter is the single-channel spectral enhancement algorithm that applies a spectral gain to the input signal. This spectral gain is computed by combining a statistical room acoustics model, minimum statistics and temporal cepstrum smoothing. We evaluated the proposed algorithm using an ASR system trained on either clean or multi-condition data, signal-based speech quality measures and listening tests. Results show that, although using multiple channels is always beneficial, the proposed algorithm is able to improve the speech quality and the performance of ASR systems, even in single-channel scenarios.

In Chapter 4 we proposed two non-intrusive speech quality measures. Both measures combine perceptually motivated features, more specifically modulation energies, and predicting functions based on machine learning. The first measure relies on a model tree that uses time-averaged features as input, while the second measure relies on an RNN that takes the time-dependency of the test signal into account by using time-varying features as input. These predicting functions have been trained using a dataset of perceptually evaluated signals that we collected for this purpose. This dataset comprises a wide range of speech enhancement algorithms, settings and acoustic scenarios. This dataset has been used to benchmark the performance of the proposed speech quality measures against both intrusive and non-intrusive measures from the literature. The performance has been evaluated when training the measures for a single category of algorithms, for several categories of algorithms, and for a mismatch between the algorithms used for training and testing. When trained and tested for a single category of algorithms, results show that the measure using an RNN as predicting function outperforms the non-intrusive measures. Furthermore, when trained and tested for several categories of algorithms, this measure outperforms both intrusive and non-intrusive benchmark measures, making it suitable for the selection of algorithms or algorithm parameters.

5.2 Further research directions

Although the contributions presented in Chapters 3 and 4 yielded promising results in terms of speech enhancement performance and speech quality prediction, many directions for further improvement could be explored.

In the field of speech enhancement, approaches based on machine learning, particularly on DNNs, are increasingly used. This is especially the case for single-channel algorithms that often rely on empirically determined parameters. Similarly as the algorithm proposed in Chapter 3, DNN-based speech enhancement algorithms also often estimate a spectral gain to be applied to the input signal. The DNNs is typically trained on large datasets for which noisy and reverberant signals as well as clean signals are available. Though such approaches have led to very promising algorithms that are able to improve speech quality and the performance of ASR systems, they introduce a new issue for algorithm development. Instead of the performance depending on empirically defined parameters, it depends on the availability of large datasets. In addition, the optimal DNN architecture is often the result of a time and resource consuming trial and error process. Since directly applying the spectral gain estimated through DNN would discard a large amount of knowledge acquired over decades of speech enhancement research, further research should ideally benefit from the advantages of both traditional speech enhancement and deep learning. This could be achieved by combining model-based algorithms, such as the one proposed in Chapter 3, with DNNs used to estimate parameters. Instead of requiring ground-truth parameters, DNNs could be trained to estimate parameters by minimizing a loss function that is related to speech quality, e.g., using a reliable speech quality measure.

Although the non-intrusive speech quality measure proposed in Chapter 4 already showed its ability to outperform state-of-the-art measures when applied to a wide range of algorithms, settings and acoustic scenarios, several extensions could be explored. First, aiming at improving the reliability of the measure, different input features and predicting functions (i.e., network architectures) could be investigated. Regardless of the chosen features and predicting function, such approaches based on machine learning need a suitable training dataset. In a controlled setting, such as considered in this thesis, labelling a large dataset of signals in term of perceived speech quality can be a daunting task. However, one may imagine using the feedback from users of, e.g., a speech communication application, to generate the required labels. Contrary to the mean opinion score used in most speech quality research, using personalized quality ratings to train the speech enhancement algorithm may lead to a personalization of the user experience.

A

STATISTICAL ANALYSIS

A repeated-measures analysis of variance (ANOVA) and post-hoc analysis was applied to the perceptual scores presented in Section 4.2 in order to assess the significance of differences between the 12 acoustic conditions. This analysis indicated that all three acoustic factors, i.e., room impulse response (RIR), noise type and signal-to-noise ratio (SNR), significantly affected the rating scores.

First, there was a significant effect of RIR ($F(2,38) = 11.4, p = 0.0013, \eta_p^2 = 0.376$) which was mainly due to significantly lower scores for RIR 2 (mean $M = 37.2$) compared to RIR 1 ($M = 42.8$) and RIR 3 ($M = 41.6$, paired $t(19) > 3.6, p < 0.002$) but no significant differences between RIR 1 and RIR 3 ($t(19) = 0.82, p = 0.42$). Second, there was a significant effect of noise type ($F(1,19) = 23.2, p < 0.001, \eta_p^2 = 0.55$) and fan noise ($M = 43.2$) was rated significantly higher compared to babble noise ($M = 37.9$). Third, there was a significant of SNR ($F(1,19) = 204.4, p < 0.001, \eta_p^2 = 0.91$) and the 5 dB condition ($M = 33.0$) was rated significantly lower compared to 15 dB condition ($M = 48.0$).

In addition, there was a significant interaction between RIR and SNR ($F(2,38) = 40.9, p < 0.001, \eta_p^2 = 0.68$). At 5 dB this was associated with significantly lower scores of RIR 1 ($M = 32.6$) and RIR 2 ($M = 28.6$) compared to RIR 3 ($M = 38.0$, paired $t(19) > 3.4, p < 0.018$, Bonferroni-corrected for multiple comparisons, n = 6) and only marginal differences between RIR 1 and RIR 2 ($t(19) = 2.8, p = 0.0624$). At 15 dB this was associated with significantly higher scores of RIR 1 ($M = 53.1$) compared to RIR 2 ($M = 45.8$) and RIR 3 ($M = 45.2$, paired $t(19) > 4.0, p < .001$) but no significant differences between RIR 2 and RIR 3 ($t(19) = 0.4, p = 0.72$).

The above two-way interaction appears to be partially driven by the significant three-way interaction of RIR, noise type, and SNR ($F(2,38) = 5.0, p = 0.0119, \eta_p^2 = 0.21$). This interaction was due to insignificant differences of RIR 1 at 5 dB fan noise ($M = 33.2$) compared to babble noise ($M = 32.0$, paired $t(19) = 0.53, p = 0.59$) and insignificant differences of RIR 3 at 15 dB fan noise ($M = 47.6$) compared to babble noise ($M = 42.9$, $t(19) = 2.2, p = 0.23$) but at least marginally significant differences between the two noise types in all other combinations of conditions ($t(19) > 2.8, p < 0.065$, using Bonferroni-correction for multiple comparisons, n=6).

B

OBSERVATIONS ON THE LSTM MEASURE

This appendix presents observation on the parameters of the LSTM network and on the ability of the LSTM measure to be applied to other datasets and to a wider range of algorithms. In order to examine if reducing the size of the input features, which could potentially reduce the computational complexity of the measure, could be beneficial, we use both modulation energy (ME) and averaged modulation energy (AME) as input features in this appendix.

When using ME, the sequence input to the long short-term memory (LSTM) network is, as in Chapter 4, a feature vector $\mathbf{e}_\ell$ of length $J \cdot B$ which for each frame is constructed as

$$\mathbf{e}_\ell = \left[\check{e}_0(0,\ell), \ldots, \check{e}_0(B-1,\ell), \ldots, \check{e}_{J-1}(0,\ell), \ldots, \check{e}_{J-1}(B-1,\ell)\right]^{\mathrm{T}}. \tag{B.1}$$

When using AME the feature coefficients in each frame are averaged, i.e., for each frame the input vector is constructed as

$$\bar{\mathbf{e}}_\ell = \left[\frac{1}{B}\sum_{b=0}^{B-1} \check{e}_0(b,\ell), \frac{1}{B}\sum_{b=0}^{B-1} \check{e}_1(b,\ell), \ldots, \frac{1}{B}\sum_{b=0}^{B-1} \check{e}_{J-1}(b,\ell)\right]^{\mathrm{T}}. \tag{B.2}$$

When using either ME or AME and as in the rest of this thesis, we use $J = 23$ channels and $B = 8$ bands.

In addition to these two types of input, in this appendix, we observe the performance of the LSTM measure in terms of ρ_{sig} and ϵ-RMSE when varying the width of each layer from 16 to 256 and the dropout used during training from 0 to 90 %.

The performance is assessed, first, on the dataset collected specifically for our work and, second, using the dataset resulting from the listening test conducted as part of the REVERB challenge.

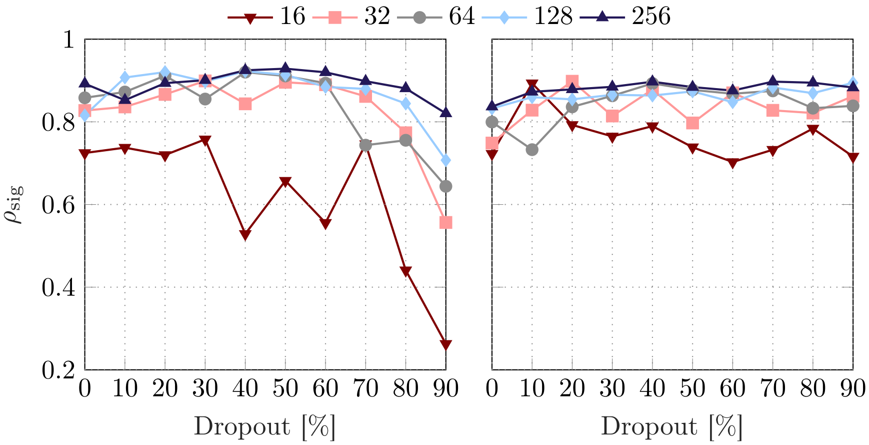

Fig. B.1: Influence of dropout and network size on obtained correlation when using either ME (left) or AME (right).

B.1 Performance on our dataset

The performance on the dataset described in Chapter 4 obtained by using either ME or AME with the considered values of dropout and network size are depicted in Fig. B.1 and in Fig. B.2. These results are obtained by training and testing the LSTM measure for all considered algorithms. Contrary to the results presented in Chapter 4, these values are obtained on a single fold.

It appears that the highest correlation among all combinations of dropout and network width is similar regardless of the choice of input, i.e., ME or AME. However, for a small sized network such as a width of 16, correlation is higher when using AME. This suggests that, unsurprisingly, smaller input has to be used if constraint such as computational power reduces the size of the network that we can use.

Encouragingly, by using AME and a small network, we can see in Fig. B.2 that both ρ_{sig} and ϵ-RMSE deteriorate only slightly. More importantly LSTM still performs better that the benchmark measures when using these settings. This suggests that, if needed, the computational complexity of the measure could be reduced and allow it to be used in realistic applications.

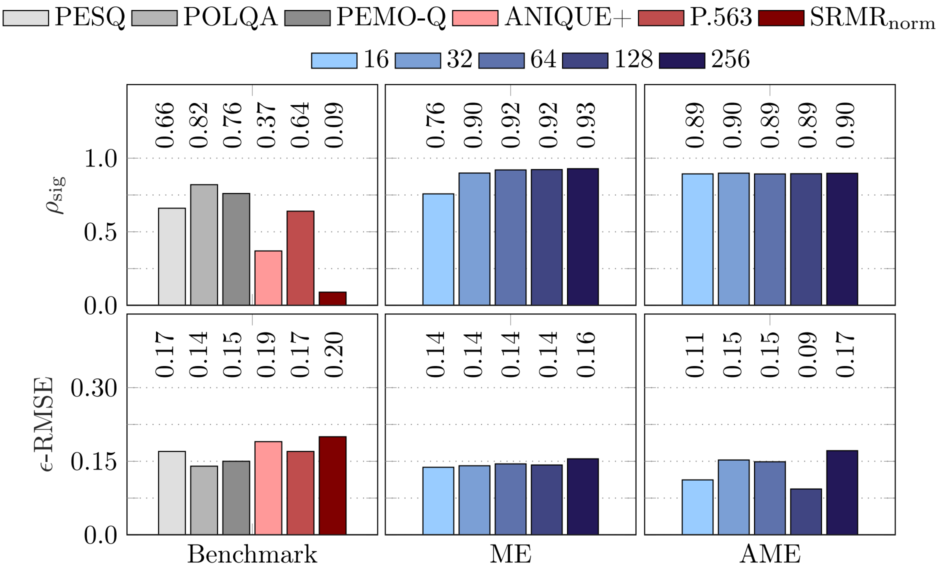

Fig. B.2: Performance obtained when using the best dropout rate for each considered network size and compared to the considered benchmark

B.2 Performance on REVERB challenge dataset

We observe the correlation of the proposed measure with the ground truth when trained using the entirety of our dataset and tested using the signals used in the REVERB challenge listening tests, whose results are reproduced in the end of Chapter 3, more specifically in Fig. 3.11 and Fig. 3.12. These correlations for the best performing dropout rate for each network width are depicted in Fig. B.3 when considering the single-channel algorithms and in Fig. B.4 when considering the multichannel algorithms.

It appears that, unsurprisingly, the correlation is lower than when training and testing using cross-validation on the same dataset. However, despite the large variety of considered algorithms, LSTM remains competitive against the benchmark non-intrusive measures. This good behavior seems, however, to be largely dependent on the network parameters. This suggests that the proposed approach is promising but that more work is needed to improve its robustness.

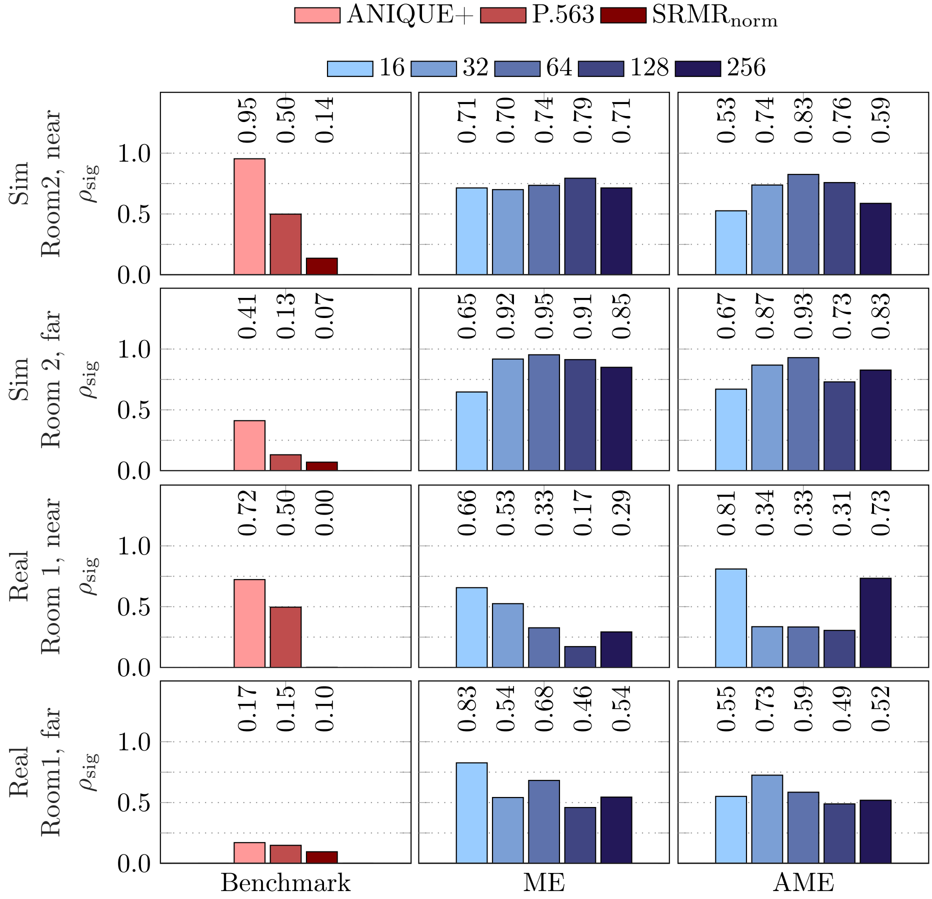

Fig. B.3: Correlation obtained using the single-channel part of the REVERB challenge dataset.

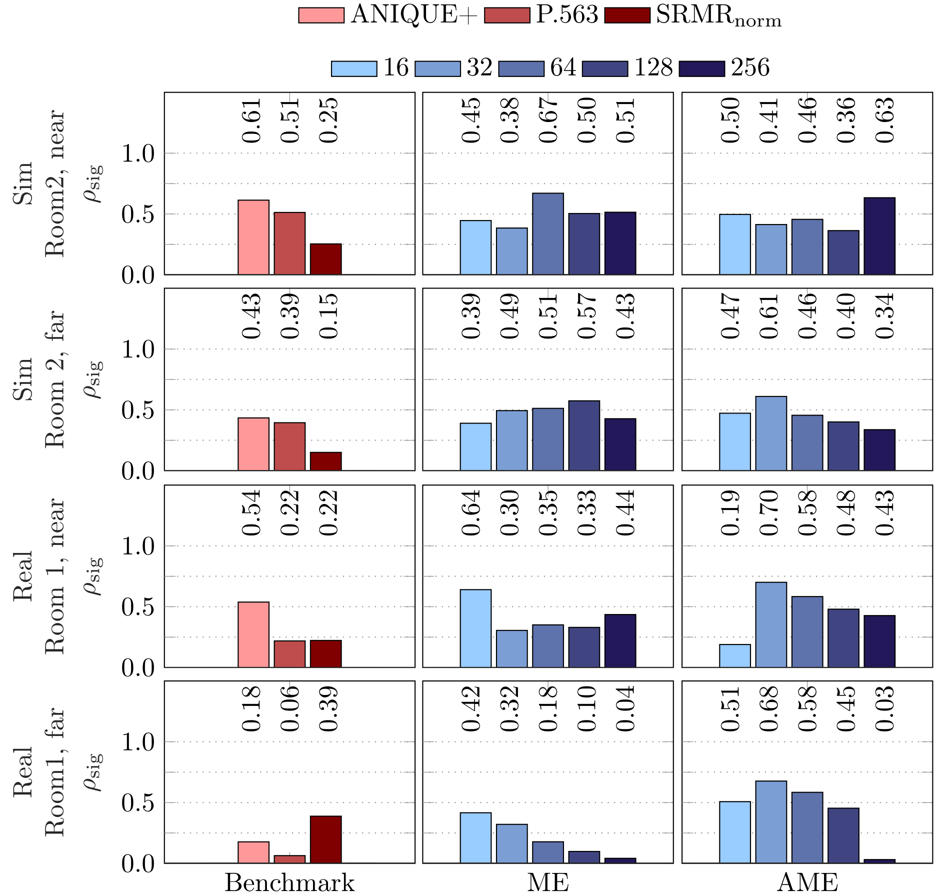

Fig. B.4: Correlation obtained using the multichannel part of the REVERB challenge dataset.

BIBLIOGRAPHY

[1] V. Pulkki and M. Karjalainen, *Communication acoustics: an introduction to speech, audio and psychoacoustics.* John Wiley & Sons, 2015.

[2] G. L. Martin, "The utility of speech input in user-computer interfaces," *International Journal of Man-Machine Studies*, vol. 30, no. 4, pp. 355–375, Apr. 1989.

[3] P. R. Cohen and S. L. Oviatt, "The role of voice input for human-machine communication," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 22, pp. 9921–9927, Oct 1995.

[4] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 75–95, Aug. 1998.

[5] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, pp. 331–342, 2006.

[6] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, and T. Nakatani, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov 2012.

[7] H. Kuttruff, *Room Acoustics*, 4th ed. London: Taylor & Francis, 2000.

[8] H. Haas, "The influence of a single echo on the audibility of speech," *Journal of the Audio Eng. Soc. (AES)*, vol. 20, no. 2, pp. 146–159, 1972.

[9] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003.

[10] I. Arweiler and M. Buchholz, J. "The influence of spectral characteristics of early reflections on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 130, no. 2, pp. 996–1005, Aug. 2011.

[11] A. Warzybok, J. Rennies, T. Brand, S. Doclo, and B. Kollmeier, "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 133, no. 1, pp. 269–282, Jan. 2013.

[12] B. Roe, D. "Deployment of human-machine dialogue systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 22, pp. 10 017–10 022, 1995.

[13] G. M. Davis, Ed., *Noise Reduction in Speech Applications.* CRC Press, May 2002.

[14] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement.* Springer-Verlag, 2005.

[15] P. C. Loizou, *Speech Enhancement: Theory and Practice.* Taylor & Francis, 2007.

[16] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing.* Springer-Verlag, 2008.

[17] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing.* Springer-Verlag, 2008.

[18] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing.* Springer-Verlag, 2009.

[19] I. Cohen, J. Benesty, and S. Gannot, *Speech processing in Modern Communication.* Springer-Verlag, 2010.

[20] R. Hendriks, T. Gerkmann, and J. Jensen, "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, Jan. 2013.

[21] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, March 2015.

[22] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[23] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement.* John Wiley & Sons, 2018.

[24] J. D. Polack, "Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2, pp. 235–244, 1993.

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

[26] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation.* Springer-Verlag, 2010.

[27] M. Miyoshi, M. Delcroix, K. Kinoshita, T. Yoshioka, T. Nakatani, and T. Hikichi, "Inverse filtering for speech dereverberation without the use of room acoustics information," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer-Verlag, 2010, ch. 9.

[28] T. Nakatani, W. Kellermann, P. Naylor, M. Miyoshi, and H. Juang, B. "Introduction to the special issue on processing reverberant speech: Methodologies and applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1673–1675, Sep. 2010.

[29] E. A. P. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer-Verlag, 2010, ch. 3.

[30] 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)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[31] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB Challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 7, pp. 1–19, Jan. 2016.

[32] A. Jukić, "Sparse multi-channel linear prediction for blind speech dereverberation," Ph.D. dissertation, University of Oldenburg, 2017.

[33] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[34] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Analysis and outcomes," *Computer Speech and Language*, vol. 46, pp. 327–344, Nov. 2017.

[35] J. P. Barker, R. Marxer, E. Vincent, and S. Watanabe, *The CHiME Challenges: Robust Speech Recognition in Everyday Environments.* Springer-Verlag, 2017, pp. 327–344.

[36] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[37] ITU-R, "Method for the subjective assessment of intermediate quality levels of coding systems," International Telecommunications Union (ITU-R), Standard BS.1534–3, Oct. 2015.

[38] P. C. Loizou, "Speech quality assessment," in *Multimedia Analysis, Processing and Communications*, W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, and H. Wang, Eds. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 623–654.

[39] S. Möller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Walermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, Oct 2011.

[40] T. H. Falk, V. Parsa, J. F. Santos, K. A, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[41] B. Cauchi, S. Goetze, and S. Doclo, "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization," in *Proc. Workshop on Speech and Multimodal Interaction in Assistive Environments (SMIAE)*, Jeju, Republic of Korea, Jul. 2012.

[42] B. Cauchi, T. Gerkmann, S. Doclo, P. A. Naylor, and S. Goetze, "Spectrally and spatially informed noise suppression using beamforming and convolutive NMF," in *Proc. Audio Eng. Soc. (AES) Conf. on Dereverberation and Rever-*

beration of Audio, Music, and Speech, Leuven, Belgium, Feb. 2016.

[43] A. Schmidt and W. Kellermann, "Informed ego-noise suppression using motor data-driven dictionaries," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 116–120.

[44] A. Subramanya, M. L. Seltzer, and A. Acero, "Automatic removal of typed keystrokes from speech signals," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 363–366, 2007.

[45] A. Sugiyama and R. Miyahara, "Tapping-noise suppression with magnitude-weighted phase-based detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[46] N. Mohammadiha and S. Doclo, "Transient noise reduction using nonnegative matrix factorization," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 27–31.

[47] D. P. Jarrett, E. A. P. Habets, and P. Naylor, *Fundamentals of Spherical Array Processing*, ser. Springer Topics in Signal Processing. Berlin Heidelberg: Springer-Verlag, 2017, vol. 9.

[48] J. Lochner and J. F. Burger, "The influence of reflections on auditorium acoustics," *J. of Sound and Vibration*, vol. 1, no. 4, pp. 426–454, 1964.

[49] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2683 –2695, Dec. 2012.

[50] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. National Academy of Sciences*, vol. 110, no. 30, pp. 12 186–12 191, 2013.

[51] A. H. Moore, M. Brookes, and P. A. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.

[52] I. Dokmanić, "Listening to distances and hearing shapes: Inverse problems in room acoustics and beyond," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2012.

[53] D. Di Carlo, "Echo-aware signal processing for audio scene analysis," Ph.D. dissertation, Université de Rennes, 2020.

[54] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, vol. 21, pp. 577–580, 1949.

[55] A. K. Nábělek and L. Robinette, "Reverberation as a parameter in clinical testing," *Audiology*, vol. 17, pp. 239–259, 1978.

[56] A. K. Nábělek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *J. of Speech and Hearing Research*, vol. 24, pp. 375–383, 1981.

[57] A. K. Nábělek and P. K. Robinson, "Monaural and binaural speech perception in reverberation for listeners of various ages," *J. of Speech and Hearing Research*, vol. 71, pp. 1242–1248, 1982.

[58] Y. Takata and A. K. Nábělek, “English consonant recognition in noise and in reverberation by Japanese and American listeners,” *J. Acoust. Soc. Am.*, vol. 88, pp. 663–666, 1990.

[59] M. R. Schroeder, “Integrated-impulse method measuring sound decay without using impulses,” *J. Acoust. Soc. Am.*, vol. 66, no. 2, pp. 497–500, 1979.

[60] D. D. Rife and J. Vanderkooy, “Transfer-function measurement with maximum-length sequences,” *Journal of the Audio Eng. Soc. (AES)*, vol. 37, no. 6, pp. 419–444, 1989.

[61] N. Ream, “Nonlinear identification using inverse-repeat m sequences,” in *Proceedings of the Institution of Electrical Engineers*, vol. 117, no. 1, 1970, pp. 213–218.

[62] N. Aoshima, “Computer-generated pulse signal applied for sound measurement,” *J. Acoust. Soc. Am.*, vol. 69, no. 5, pp. 1484–1488, May 1981.

[63] A. J. Berkhout, M. M. Boone, and C. Kesselman, “Acoustic impulse response measurement: A new technique,” *Journal of the Audio Eng. Soc. (AES)*, vol. 32, no. 10, pp. 740–746, Oct. 1984.

[64] 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.

[65] G. W. Elko, E. Diethorn, and T. Gänsler, “Room impulse response variation due to thermal fluctuation and its impact on acoustic echo cancellation,” in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Kyoto, Japan, Sep. 2003.

[66] J. A. Cadzow, “Blind deconvolution via cumulant extrema,” *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 24–42, May 1996.

[67] W. Xue, M. Brookes, and A. Naylor, P. “Cross-correlation based under-modelled multichannel blind acoustic system identification with sparsity regularization,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, 2016.

[68] G. Xu, H. Liu, L. Tong, and T. Kailath, “A least-squares approach to blind channel identification,” *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[69] K. Abed-Meraim, W. Qiu, and Y. Hua, “Blind system identification,” *Proc. IEEE*, vol. 85, no. 8, pp. 1310–1322, 1997.

[70] M. Wu, “Cross-relation based blind identification of SIMO acoustic systems and applications,” Ph.D. dissertation, Imperial College London, 2017.

[71] I. Kodrasi, “Dereverberation and noise reduction techniques based on acoustic multi-channel equalization,” Ph.D. dissertation, University of Oldenburg, 2015.

[72] F. Lim, “Robust multichannel equalization for blind speech dereverberation,” Ph.D. dissertation, Imperial College London, 2016.

[73] B. Cauchi, P. A. Naylor, T. Gerkmann, S. Doclo, and S. Goetze, “Late reverberant spectral variance estimation using acoustic channel equalization,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, Sep. 2015.

[74] S. Goetze, M. Kalinger, A. Mertins, and K. D. Kammeyer, "System identification for multi-channel listening-room compensation using an acoustic echo canceller," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, May 2008, pp. 224–227.

[75] J. Mourjopoulos and M. Paraskevas, "Pole and zero modeling of room transfer functions," *J. of Sound and Vibration*, vol. 146, no. 2, pp. 281–302, 1991.

[76] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 320–328, 1994.

[77] G. Vairetti, E. De Sena, M. Catrysse, S. Jensen, M. Moonen, and T. van Watershoot, "A scalable algorithm for physically motivated and sparse approximation of room impulse responses with orthonormal basis functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1547–1561, Jul. 2017.

[78] M. R. Schroeder, "On frequency response curves in rooms: Comparison of experimental, theoretical and Monte-Carlo results for the average frequency spacing between maxima," *J. Acoust. Soc. Am.*, vol. 34, no. 1, pp. 76–80, 1962.

[79] 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.

[80] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 255–267, Feb. 2019.

[81] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–774, Sep. 2009.

[82] A. Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[83] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, ser. SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag, 2012.

[84] B. Yang, "A study of inverse short-time Fourier transform," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 3541–3544.

[85] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, "A brief survey of speech enhancement," in *The Electronic Handbook*, 2nd ed. CRC Press, Feb. 2005.

[86] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer-Verlag, 2008, ch. 44.

[87] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[88] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[89] J. M. G. J. Ramirez and J. C. Segura, “Voice activity detection. fundamentals and speech recognition system robustness,” in *Noise Reduction in Speech Applications*, M. Grimm and K. Kroschel, Eds. InTech, 2007, ch. 1, pp. 1–22.

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

[91] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[92] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech Communication*, vol. 48, no. 2, pp. 220–231, Feb. 2006.

[93] 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)*, Dallas, Texas, USA, Mar. 2010, pp. 4266–4269.

[94] 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.

[95] H. Attias and L. Deng, “Speech denoising and dereverberation using probabilistic models,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, pp. 758–764, 2001.

[96] C. S. J. Doire, M. Brookes, P. A. Naylor, C. M. Hicks, D. Betts, M. A. Dmour, and S. H. Jensen, “Single-channel online enhancement of speech corrupted by reverberation and noise,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 572–587, Mar. 2017.

[97] J. Porter and S. Boll, “Optimal estimators for spectral restoration of noisy speech,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 9, Mar. 1984, pp. 53–56.

[98] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[99] J. S. Erkelens, R. C. Hendriks, R. . Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[100] M. Sun, Y. Li, F. Gemmeke, J. and X. Zhang, “Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[101] N. Mohammadiha and S. Doclo, “Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 2, pp. 276–289, Feb. 2016.

[102] M. Kolbæk, H. Tan, Z. and J. Jensen, “Speech intelligibility potential of general and specialized deep neural network based speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 149–163, Jan. 2017.

[103] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[104] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[105] T. May and T. Gerkmann, “Generalization of supervised learning for binary mask estimation,” in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Juan-les-Pins, France, 2014, pp. 154–158.

[106] F. Xiong, B. T. Meyer, B. Cauchi, A. Jukić, S. Doclo, and S. Goetze, “Performance comparison of real-time single-channel speech dereverberation algorithms,” in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, 2017, pp. 126–130.

[107] J. Chen and D. Wang, “DNN based mask estimation for supervised speech separation,” in *Audio Source Separation*, S. Makino, Ed. Springer-Verlag, 2018, pp. 207–235.

[108] H. Cox, R. M. Zeskind, and T. Kooij, “Practical supergain,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 3, pp. 393–398, 1986.

[109] B. D. van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[110] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques.* Prentice-Hall, 1993.

[111] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[112] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction,” *Speech Communication*, vol. 49, no. 7–8, pp. 636–656, Aug. 2007.

[113] J. Bitzer, K.-D. Kammeyer, and K. U. Simmer, “An alternative implementation of the superdirective beamformer,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 1999.

[114] O. L. Frost, III, “An algorithm for linearly constrained adaptive array processing,” *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[115] J. Capon, “High resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[116] 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.

[117] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[118] N. Madhu, "Acoustic source localization: Algorithms, applications and extensions to source separation," Ph.D. Thesis, Ruhr-Universität Bochum, May 2009.

[119] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications.* Springer-Verlag, 2001.

[120] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC)*, Darmstadt, Germany, Sep. 2001, pp. 31–34.

[121] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[122] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using maxNSR blocking matrix," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1493–1508, Sep. 2015.

[123] M. Taseska and E. A. P. Habets, "Non-stationary noise PSD matrix estimation for multi-channel blind speech extraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 11, pp. 2223–2236, Nov. 2017.

[124] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shangai, China, 2016, pp. 151–155.

[125] I. Kodrasi and S. Doclo, "Joint late reverberation and noise power spectral density estimation in a spatially homogeneous noise field," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB,Canada, Apr. 2018, pp. 441–445.

[126] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1106–1118, Jun. 2018.

[127] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 1981–1985.

[128] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech and Language*, vol. 46, pp. 374–385, May 2017.

[129] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise de-reverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[130] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Dereverberation by using time-variant nature of speech production system," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 2, Aug. 2007.

[131] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 17, pp. 534–545, May 2009.

[132] T. Nakatani, K. Yoshioka, K. Kinoshita, M. Miyoshi, and H. Juang, B. "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[133] T. Yoshioka and T. Nakatani, "Dereverberation for reverberation robust microphone arrays," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakesh, Morocco, Sep. 2013.

[134] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Group sparsity for MIMO speech dereverberation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015.

[135] T. Dietzen, A. Spriet, W. Tirry, M. Doclo, S. Moonen, and T. van Waterschoot, "Partitioned blocked frequency domain Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Xián, China, Sep. 2016.

[136] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 101–105, Jan. 2017.

[137] T. Xiang, J. Lu, and K. Chen, "Multi-channel adaptive dereverberation robust to abrupt change of target speaker position," *IEEE Signal Processing Letters*, vol. 145, no. 3, pp. 250–256, Mar. 2019.

[138] M. S. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6, 1998, pp. 3613–3616.

[139] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[140] T. Yoshioka, T. Nakatani, K. Kinoshita, and M. Miyoshi, "Speech dereverberation and denoising based on time varying speech model and autoregressive reverberation model," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer-Verlag, Jan. 2010, ch. 6.

[141] M. Togami and Y. Kawaguchi, "Noise robust speech dereverberation with Kalman smoother," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 7447–7451.

[142] Audio Software Engineering and Siri Speech Team, "Optimizing Siri on HomePod in far-field settings," *Apple Machine Learning Journal*, vol. 1, no. 12, 2018.

[143] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home

assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, Oct 2019.

[144] Y. Wang, "Model-based approaches to robust speech recognition in diverse environments," Ph.D. dissertation, Cambridge University, 2017.

[145] M. Johnson, S. Lapkin, V. Long, P. Sanchez, H. Suominen, J. Basilakis, and L. Dawson, "A systematic review of speech recognition technology in health care," *BMC medical informatics and decision making*, vol. 14, no. 1, p. 94, 2014.

[146] K. Zinchenko, C. Wu, and K. Song, "A study on speech recognition control for surgical robot," *IEEE Trans. Industrial Informatics*, vol. 13, no. 2, pp. 607–615, Apr. 2017.

[147] E. Yilmaz, V. Mitra, G. Sivaraman, and H. Franco, "Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech," *Computer Speech and Language*, vol. 58, pp. 319–334, 2019.

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

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

[150] N. Moritz, "Amplitude modulation analysis for automatic speech recognition," Ph.D. dissertation, University of Oldenburg, 2016.

[151] N. Sainath, T. J. Weiss R. W. Wilson, K. B. Li, A. Narayanan, E. Variana, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965–979, 2017.

[152] T. Ochiai, S. Watanabe, H. Takaaki, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. Intl. Conf. Machine Learning (ICML)*, Stockholm, Sweden, 2017.

[153] B. Li, S. Chang, N. Sainath, T. R. Pang, Y. He, T. Strohman, and Y. Wu, "Towards fast and accurate streaming end-to-end ASR," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6069–6073.

[154] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2801–2810, Jun. 2002.

[155] C. S. J. Doire, M. Brookes, and P. A. Naylor, "Robust and efficient Bayesian adaptive psychometric function estimation," *J. Acoust. Soc. Am.*, vol. 141, no. 4, pp. 2501–2512, 2017.

[156] B. Kollmeier, A. Warzybok, S. Hochmuth, M. Zokoll, N. Uslar, V. T. Brand, and K. Wagener, "The multilingual matrix test: Principles, applications, and comparison across languages: A review," *International journal of audiology*, vol. 54, pp. 1–14, Sep. 2015.

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

[158] *Subjective assessment of sound quality*, International Telecommunications Union (ITU-R) Recommendation BS. 562–3, 1990.

[159] E. Rothauser, G. Urbanke, and W. Pachl, "A comparison of preference measurement methods," *J. Acoust. Soc. Am.*, vol. 49(4), pp. 1297–1308, 1970.

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

[161] W. Voiers, A. Sharpley, and I. Panzer, "Evaluating the effects of noise on voice communication systems," in *Noise Reduction in Speech Applications*, G. Davis, Ed. CRC Press, May 2002.

[162] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Mass, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka. (2015, Sep.) Summary of the REVERB Challenge. http://reverb2014.dereverberation.com/workshop/slides/reverb_summary.pdf.

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

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

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

[166] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[167] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 6–16, Jan. 2014.

[168] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, vol. 7, Sydney, Australia, 1998, pp. 2819–2822.

[169] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Tulsa, OK, USA, Apr. 1978, pp. 586–590.

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

[171] M. Sambur and N. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 6, pp. 488–494, 1976.

[172] A. H. Gray, Jr. and J. D. Markel, “Distance measures for speech processing,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 380–391, Oct. 1976.

[173] I. Smith, J. O. and J. S. Abel, “Bark and ERB bilinear transforms,” *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, 1999.

[174] S. Wang, A. Sekey, and A. Gersho, “An objective measure for predicting subjective quality of speech coders,” *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, Jun. 1992.

[175] W. Yang, M. Benbouchta, and R. Yantorno, “Performance of the modified Bark spectral distortion as an objective speech quality measure,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1998, pp. 541–544.

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

[177] ITU-T, “Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals,” International Telecommunications Union (ITU-T), Standard P.863, Jan. 2011.

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

[179] J. Friedman, “Multivariate adaptive regression splines,” *Ann. Stat.*, vol. 19, no. 1, pp. 1–67, 1991.

[180] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, “A non-intrusive short-time objective intelligibility measure,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5085–5089.

[181] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, and M. G. Christensen, “Non-intrusive codebook-based intelligibility prediction,” *Speech Communication*, vol. 101, pp. 85–93, 2018.

[182] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, “Predicting speech intelligibility with deep neural networks,” *Computer Speech and Language*, vol. 48, pp. 51–66, 2018.

[183] R. Schädler, M. D. Hülsmeier, A. Warzybok, and B. Kollmeier, “Individual aided speech-recognition performance and predictions of benefit for listeners with impaired hearing employing FADE,” *Trends in Hearing*, vol. 24, 2020.

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

[185] 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.

[186] 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)*, Antibes, France, Sep. 2014, pp. 55–59.

[187] J. Santos and T. Falk, "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users," vol. 22, no. 12, Dec. 2014, pp. 2197–2206.

[188] S. Goetze, A. Warzybok, I. Kodrasi, J. Jungmann, B. Cauchi, J. Rennies, E. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Antibes, France, Sep. 2014, pp. 233–237.

[189] A. Avila, B. Cauchi, S. Goetze, S. Doclo, and T. Falk, "Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Xiánn, China, Sep. 2016, pp. 1–5.

[190] I. Kodrasi, B. Cauchi, S. Goetze, and S. Doclo, "Instrumental and perceptual evaluation of dereverberation techniques based on robust acoustic multichannel equalization," *Journal of the Audio Eng. Soc. (AES)*, vol. 65, no. 1/2, pp. 117–129, 2017.

[191] M. Karbasi, A. Abdelhaziz, H. Meutzner, and D. Kolossa, "Blind non-intrusive speech intelligibility prediction using twin-HMMs," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 625–629.

[192] D. S. Kim and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, pp. 221–236, 2007.

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

[194] D. Sharma, Y. Wang, P. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, pp. 84–94, Jun. 2016.

[195] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech-enhancement scheme," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[196] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 61, pp. 1–12, Jul. 2015.

[197] B. Cauchi, J. Santos, K. Siedenburg, T. Falk, P. Naylor, S. Doclo, and S. Goetze, "Predicting the quality of processed speech by combining

modulation-based features and model trees," in *Proc. ITG Conf. on Speech Communication*, Paderborn, Germany, Oct. 2016.

[198] B. Cauchi, K. Siedenburg, J. Santos, T. Falk, S. Doclo, and S. Goetze, "Non-intrusive speech quality prediction using modulation energies and LSTM-network," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1151–1163, Jul. 2019.

[199] M. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 3, pp. 374–390, Jun. 1981.

[200] D. Pearce and H. G. Hirsch, "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.

[201] F. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*, Y. Deville, S. Gannot, R. Mason, M. D. Plumbley, and D. Ward, Eds. Springer-Verlag, 2018, pp. 293–305.

[202] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, Michigan, U.S.A, May 1995, pp. 81–84.

[203] "Documentation about the room impulse responses and noise data used for the REVERB Challenge SimData," [Online] Available: http://reverb2014.dereverberation.com/tools/Document_RIR_noise_recording.pdf.

[204] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multichannel Wall Street Journal audio-visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE Workshop Autom. Speech Recognition and Understanding (ASRU)*, Cancún, Mexico, Dec. 2005, pp. 357–362.

[205] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, 3rd ed. Cambridge University Engineering Dept, Dec. 2009.

[206] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[207] R. Maas, E. A. P. Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 297–300.

[208] E. A. P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. IV, Honolulu, USA, Apr. 2007, pp. 901–904.

[209] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets, "An informed spatial filter for dereverberation in the spherical harmonic domain," in *Proc. IEEE*

Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013, pp. 669–673.

[210] A. Schwarz, K. Reindl, and W. Kellermann, "On blocking matrix-based dereverberation for automatic speech recognition," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sept 2012.

[211] A. Schwarz, K. Reindl, and W. Kellermann, "A two-channel reverberation suppression scheme based on blind signal separation and wiener filtering," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, May 2012, pp. 113–116.

[212] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[213] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Communication*, vol. 34, pp. 3–12, 2001.

[214] H. W. Löllmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2009, pp. 3989–3992.

[215] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2008, pp. 4037–4040.

[216] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2008, pp. 4897–4900.

[217] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[218] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[219] 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.

[220] I. Gradshteyn and I. Ryzhik, *Table of integrals, series, and products.* Boston: Academic Press, Inc., 1994.

[221] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[222] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*,

vol. 32, no. 200, pp. 675–701, 1937.

[223] J. Gibbons and S. Chakraborti, *Nonparametric statistical inference.* Springer-Verlag, 2011.

[224] K. Ohtani, T. Komatsu, T. Nishino, and K. Takeda, "Adaptive dereverberation method based on complementary Wiener filter and modulation transfer function," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[225] K. Kondo, "A computationally restrained and single-channel blind dereverberation method utilizing iterative spectral modifications," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[226] X. Xiao, Z. Shengkui, H. H. Nguyen, D. Z. Xionghu, D. Jones, S. Chang, E. and H. Li, "The NTU-ADSC systems for reverberation challenge 2014," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[227] S. Gonzalez and M. Brookes, "Mask-based enhancement for very low quality speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, May 2014.

[228] J. Veras, T. Prego, A. Lima, T. Ferreira, and S. Netto, "Speech quality enhancement based on spectral subtraction," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[229] G. Lopez, G. Richard, Y. Grenier, and I. Bourmeyster, "Reverberation suppression based on sparse linear prediction in noisy environments," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[230] M. Moshirynia, F. Razzaza, and A. Haghbin, "A speech dereverberation method using adaptive sparse dictionary learning," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[231] R. Leng, Y J. Dennis, Z. T. Ng, W. and H. Dat, T. "PBF-GSC beamforming for ASR and speech enhancement," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[232] M. Yu and F. Soong, "Speech dereverberation by constrained and regularized multi-channel spectral decomposition: evaluated on reverb challenge," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[233] N. Epain, T. Noohi, and C. Jin, "Sparse recovery method for dereverberation," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[234] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and

recognition technologies for the REVERB challenge," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[235] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," International Telecommunications Union (ITU-T), Standard P.835, Nov. 2003.

[236] P. Flipsen, "Measuring the intelligibility of conversational speech in children," *Clinical Linguistics and Phonetics*, vol. 20, no. 4, pp. 303–312, 2006.

[237] R. Huber, J. Ooster, and T. Meyer, B. "Single-ended speech quality prediction based on automatic speech recognition," *Journal of the Audio Eng. Soc. (AES)*, vol. 66, no. 10, pp. 759–769, Oct. 2018.

[238] 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.

[239] E. Frank, Y. Wang, S. Ingliss, G. Holmes, and I. H. Witten, "Using model trees for classification," *Machine Learning*, vol. 32, no. 1, pp. 63–76, 1998.

[240] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.

[241] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Intl. Conf. Machine Learning (ICML)*, 2013, pp. 1310–1318.

[242] J. Santos and T. Falk, "Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics," in *Proc. Conf. Audio Eng. Soc. (AES)*, Feb. 2016.

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

[244] ITU-T, *General Performance Objectives Applicable to All Modern International Circuits and National Extension Circuits*, International Telecommunications Union (ITU-T) Recommendation G.151, Mar. 1980.

[245] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[246] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[247] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[248] C. Sprinthall, R. and T. Fisk, S. *Basic statistical analysis.* Prentice Hall Englewood Cliffs, NJ, USA, 2013.

[249] ITU-T, "Statistical evaluation procedure for POLQA," International Telecommunications Union (ITU-T), Standard TD 12rev1 (WP 2/12), Mar. 2009.

LIST OF PUBLICATIONS BY THE AUTHOR

Peer-reviewed Journal Papers

[J3] B. Cauchi, K. Siedenburg, J. Santos, T. Falk, S. Doclo, and S. Goetze, “Non-intrusive speech quality prediction using modulation energies and LSTM-network,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1151-1163, Jul. 2019.

[J2] I. Kodrasi, B. Cauchi, S. Goetze, and S. Doclo, “Instrumental and perceptual evaluation of dereverberation techniques based on robust acoustic multichannel equalization,” *Journal of the Audio Eng. Soc. (AES)*, vol. 65, no. 1/2, pp. 117-129, 2017.

[J1] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, “Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech,” *EURASIP Journal on Advances in Signal Processing*, vol. 61, pp. 1-12, Jul. 2015.

Peer-reviewed Conference Papers

[C10] F. Xiong, B. T. Meyer, B. Cauchi, A. Jukić, S. Doclo, and S. Goetze, “Performance comparison of real-time single-channel speech dereverberation algorithms,” in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, 2017, pp. 126-130.

[C9] B. Cauchi, J. Santos, K. Siedenburg, T. Falk, P. Naylor, S. Doclo, and S. Goetze, “Predicting the quality of processed speech by combining modulation-based features and model trees,” in *Proc. ITG Conf. on Speech Communication*, Paderborn, Germany, Oct. 2016.

[C8] A. Avila, B. Cauchi, S. Goetze, S. Doclo, and T. Falk, “Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech,” in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Xián, China, Sep. 2016, pp. 1-5.

[C7] B. Cauchi, H. Javed, T. Gerkmann, S. Doclo, S. Goetze, and P. Naylor, “Perceptual and instrumental evaluation of the perceived level of reverberation,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shangai, China, March 2016, pp. 629-633.

[C6] B. Cauchi, T. Gerkmann, S. Doclo, P. A. Naylor, and S. Goetze, “Spectrally and spatially informed noise suppression using beamforming and convolutive

NMF," in *Proc. Audio Eng. Soc. (AES) Conf. on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, Feb. 2016.

[C5] B. Cauchi, P. A. Naylor, T. Gerkmann, S. Doclo, and S. Goetze, "Late reverberant spectral variance estimation using acoustic channel equalization," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, Sep. 2015.

[C4] S. Goetze, A. Warzybok, I. Kodrasi, J. Jungmann, B. Cauchi, J. Rennies, E. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Antibes, France, Sep. 2014, pp. 233-237.

[C3] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech-enhancement scheme," in *Proc. REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge Workshop*, Florence, Italy, May 2014.

[C2] J. Schröder, N. Moritz, M. R. Schädler, B. Cauchi, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "On the use of spectro-temporal features for the IEEE AASP challenge detection and classification of acoustic scenes and events," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[C1] B. Cauchi, S. Goetze, and S. Doclo, "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization," in *Proc. Workshop on Speech and Multimodal Interaction in Assistive Environments (SMIAE),*, Jeju, Republic of Korea, Jul. 2012.