Single-Microphone Speech Dereverberation based on Multiple-Step Linear Predictive Inverse Filtering and Spectral Subtraction

Ali Baghaki

A Thesis

in
The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements for the Degree of

Master of Applied Science (Electrical and Computer Engineering) at

Concordia University,

Montreal, Quebec, Canada

August 2013

© Ali Baghaki, 2013

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Ali Baghaki

Entitled: “Single-Microphone Speech Dereverberation based on Multiple-Step Linear Predictive Inverse Filtering and Spectral Subtraction”

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science

Complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

____ Chair
Dr. M. Z. Kabir

____ Examiner, External
Dr. R. Bhat (MIE) To the Program

____ Examiner
Dr. S. Hashtrudi Zad

____ Supervisor
Dr. M. O. Ahmad

____ Supervisor
Dr. M. N. S. Swamy

Approved by: ____
Dr. W. E. Lynch, Chair
Department of Electrical and Computer Engineering

____________20_____ __________________________________
Dr. C. W. Trueman
Interim Dean, Faculty of Engineering
and Computer Science

ABSTRACT

Single-Microphone Speech Dereverberation based on Multiple-Step
Linear Predictive Inverse Filtering and Spectral Subtraction

Ali Baghaki

Single-channel speech dereverberation is a challenging problem of deconvolution of reverberation, produced by the room impulse response, from the speech signal, when only one observation of the reverberant signal (one microphone) is available. Although reverberation in mild levels is helpful in perceiving the speech (or any audio) signal, the adverse effect of reverberation, particularly at high levels, could both deteriorate the performance of automatic recognition systems and make it less intelligible by humans. Single-microphone speech dereverberation is more challenging than multi-microphone speech dereverberation, since it does not allow for spatial processing of different observations of the signal.

A review of the recent single-channel dereverberation techniques reveals that, those based on LP-residual enhancement are the most promising ones. On the other hand, spectral subtraction has also been effectively used for dereverberation particularly when long reflections are involved. By using LP-residuals and spectral subtraction as two promising tools for dereverberation, a new dereverberation technique is proposed. The first stage of the proposed technique consists of pre-whitening followed by a delayed long-term LP filtering whose kurtosis or skewness of LP-residuals is maximized to control the weight updates of the inverse filter. The second stage consists of nonlinear spectral subtraction. The proposed two-stage dereverberation scheme leads to two separate algorithms depending on whether kurtosis or skewness maximization is used to establish a feedback function for the weight updates of the adaptive inverse filter.

It is shown that the proposed algorithms have several advantages over the existing major single-microphone methods, including a reduction in both early and late reverberations, speech enhancement even in the case of very high reverberation time, robustness to additive background noise, and introducing only a few minor artifacts. Equalized room impulse responses by the proposed algorithms have less reverberation times. This means the inverse-filtering by the proposed algorithms is more successful in dereverberating the speech signal. For short, medium and high reverberation times, the signal-to-reverberation ratio of the proposed technique is significantly higher than that of the existing major algorithms. The waveforms and spectrograms of the inverse-filtered and fully-processed signals indicate the superiority of the proposed algorithms. Assessment of the overall quality of the processed speech signals by automatic speech recognition and perceptual evaluation of speech quality test also confirms that in most cases the proposed technique yields higher scores and in the cases that it does not do so, the difference is not as significant as the other aspects of the performance evaluation. Finally, the robustness of the proposed algorithms against the background noise is investigated and compared to that of the benchmark algorithms, which shows that the proposed algorithms are capable of maintaining a rather stable performance for contaminated speech signals with SNR levels as low as 0 dB.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my thesis advisors, Professor M.N.S. Swamy and Professor M. Omair Ahmad, for all their trust, their advice, their patience and their financial support. Indeed, it has been an honor for me to do research towards my masters' degree with their supervision. As well, I should acknowledge Dr. Peter Kabal from McGill University for the main initiative of this thesis was his proposed idea for his course project of DSP II.

During the period of my stay in the signal processing lab EV 10.119 I have also enjoyed the company and the help of my colleagues and friends. I am especially grateful to Mufleh Al-Shatnawi, Sarath Somasekharan Pillai and Yaser Mohammad-Taheri for their assistance and friendship.

Life as a graduate student is not easy. Even so, I was lucky to meet some great people who inspired me and helped me to overcome the difficulties. I thank them all. I am especially grateful to Pouya Jabbari for his support and for being a great friend.

Last but not least, I owe more than thanks to my family members, which include my parents and my sister in Iran. Without them, I would not be able to succeed throughout my life. I am always grateful to them for their encouragement and financial and mental support.

“Essentially, all models are wrong, but some are useful.”

– George E. P. Box

To My Loving Family

Contents

List of Figures

List of Tables

List of Symbols

$A(z)$	The shaping filter in the human speech production system
$\hat{a}(l)$	Linear prediction filter coefficients
$B(z)$	Transfer function of the acoustic channel from speaker to microphone
D	The delay number of the delayed long-term linear prediction
$E\{.\}$	Expectation operator
$E_z(i)$	Energy value of the inverse-filtered speech at the time frame i
$E_p(i)$	Energy value of the processed speech signal at the time frame i
$\tilde{e}(n)$	Linear prediction residual (error)
$\hat{e}(n)$	Enhanced linear prediction residual
$f_1(t)$	Feedback function for the weight update of the inverse filter in kurtosis maximization
$f_2(t)$	Feedback function for the weight update of the inverse filter in skewness maximization
f_s	The sampling frequency
$g(n)$	Impulse response of the filter combining human speech production system and the effect of the room impulse response
$H(n)$	Representation of $h(n)$ for frequency-block implementation
$h(n)$	The inverse filter in time domain

$j_1(t)$	Kurtosis function
$j_2(t)$	Skewness function
L_{z_d}	The Bark spectrum of the direct signal
$L_{\hat{z}_d}$	The Bark spectrum of the enhanced signal
m	Frame number (unless otherwise specified)
N	Number of filter taps in the delayed long-term linear prediction
$\hat{s}(n)$	Dereverberated speech signal
$S_l(k;i)$	Short-term power spectrum of the late impulse components
$S_p(k;i)$	The short-term power spectrum of the processed speech
$S_z(k;i)$	Short-term power spectrum of the inverse-filtered speech at frequency bin k and frame i
$u(n)$	Excitation signal in the human speech production system
$w(n)$	The filter coefficients of the delayed long-term linear prediction
$x_r(t)$	Linear prediction residual signal of multiple-step linear predictor
$\hat{x}_r(t)$	A frame of $x_r(t)$
$z(t)$	The processed speech
$\tilde{z}(t)$	The inverse-filtered speech
α	The overall spread of the Rayleigh smoothing function
β	Parameter controlling the smoothness of moment estimates in kurtosis and skewness maximization

γ	Scaling factor controlling the relative strength of the late impulse components
σ_u^2	Variance of the excitation signal, $u(n)$
ε	The threshold of attenuation of late impulse components
μ	Parameter adjusting the learning rate for the weight updates of the inverse filter
v_1	The first threshold for silent detection
v_2	The second threshold for silent detection
$\omega(i)$	The Rayleigh smoothing function at the time frame i
$\|\|.\|\|$	Norm operator

List of Abbreviations

AIR	Acoustic impulse response
AR	Auto regressive
ASR	Automatic speech recognition
BSD	Bark spectral distortion
CMS	Cepstral mean subtraction
dB	Decibel
DLLP	Delayed long-term linear prediction
DOA	Direction of arrival
DRR	Direct to reverberation ratio
EDC	Energy decay curve
EMBSD	Enhanced modified Bark spectral distortion
FFT	Fast fourier transform
FIR	Finite impulse response
IIR	Infinite impulse response
LMS	Least mean square

LP	Linear prediction
LPC	Linear prediction coefficient
MBSD	Modified Bark spectral distortion
MCLT	Modulated complex lapped transform
MMSE	Minimum mean square error
MOS	Mean opinion score
MTF	Modulation transfer function
NsegSRR	Normalized segmental signal to reverberation ratio
PDA	Personal digital assistant
PDF	Probabilistic density functions
PESQ	Perceptual evaluation of speech quality
POLQA	Perceptual objective listening quality assessment
PSD	Power spectral density
RIR	Room impulse response
SNR	Signal to noise ratio
SRR	Signal to reverberation ratio

STFT	Short time Fourier Transform
STPSD	Short time power spectral density
VOIP	Voice over internet protocol
WER	Word error rate

Chapter 1

Introduction

1.1. Background

The phenomenon of reverberation has been known to humankind since prehistoric era when people were residing in caves. According to sources, the footprint of some understanding of the reverberation phenomenon can be found in prehistoric cave art [1]. In Plato's *Republic*, there is reference to the reflected speech from the walls, implying a comprehension of reverberation. Initial scientific study of reverberation dates back to the mid-to-late 20th century by pioneers such as Bolt [2] and Haas [3].

There is no doubt in the fact that reverberation is a useful phenomenon in everyday life. For example, by taking advantage of the two ears, speech intelligibility is enhanced by spatial processing in the human hearing system. This gives the humans the capability to some degree of source separation in perceiving mixed sounds [1]. As another example, in music audio processing, stereo or surround sound reproduction enhances the realism and joy of the recorded music. Therefore, the question that comes to the mind is: "*As reverberation is present in everyday life experience as a useful phenomenon, why should one be interested in removing reverberation from speech using dereverberation processing*?". The short answer to this question is that usefulness or harmfulness of reverberation is application-dependant [1]. The demand for high-quality hands-free speech input is constantly increasing. This is due to the growing use of portable devices such as mobile telephones, personal digital assistant (PDA) devices and laptop computers equipped with voice over internet protocol (VoIP). In addition, the broadband internet access is constantly growing worldwide. As a result, several advanced speech applications such as wideband teleconferencing with automatic camera steering, automatic speech-to-text conversion, speaker identification, voice-controlled device operation and car interior communication systems, have appeared. Hearing aids is another application in which the quality of the speech by a distant talker is important [1]. In all these examples, the desired acoustical source might be located at a distance from the microphone.

As depicted in Fig. 1.1, the desired source produces sound waves. In addition to the direct sound wave travelling the direct path between the source and the microphone, parts of the energy of the source signal reaches the microphone only after being scattered and reflected from walls, floor, ceiling and other surfaces. This phenomenon is called reverberation. As a result, in general, the resulting direct signal might be degraded by reverberation, background noise, and other interferences [4].

One of the degradations in the desired signal occurs when a signal is recorded in an

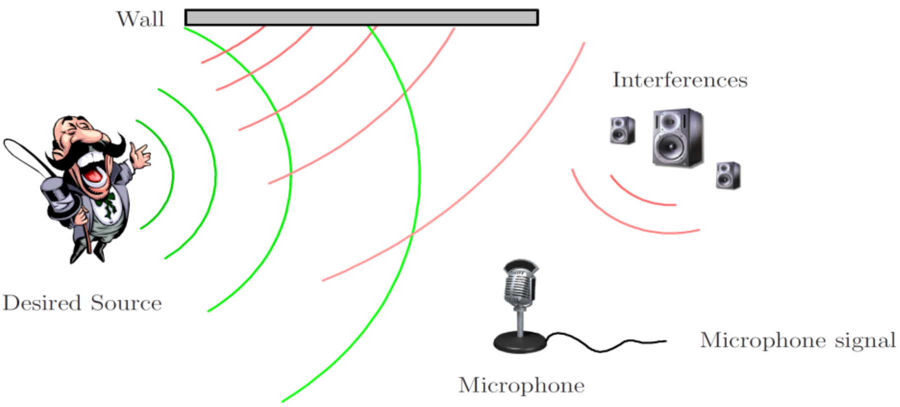


Fig. 1.1. Illustration of a desired source, a microphone, and interfering sources [4].

enclosed space, e.g., an office room or a living room and thus is affected by the acoustic channel. The received microphone signals are typically degraded by two factors: (i) reflections by the multi-path propagation of the sound to the microphone(s) and (ii) noise produced by interfering sources. This happens more severely when the microphone(s) are not located near the desired source [1], [4].

It should be noted that many, if not all, existing acoustic signal processing techniques, e.g. existing source localization and source separation techniques, end up in a complete failure or a drastically reduced performance in the presence of reverberation. Nowadays, while state-of-the-art acoustic signal processing algorithms are available for noise suppression, the development of efficient and practical algorithms that can reduce the reverberation is still a major challenge.

The key difference between noise and reverberation is that the degradation produced by reverberation is dependent on the desired signal, whereas that of noise can be assumed to be independent of the desired signal [1], [4].

The harmful perceptual effects of reverberation generally increase with increasing distance between the source and the microphone. Besides, since reflections arrive at the microphone at different times, reverberation causes blurring of speech phonemes. These damaging effects can severely deteriorate the intelligibility, the performance of voice-controlled systems, and the performance of speech coding algorithms used in telephone systems. Hence, reducing these harmful effects is evidently of substantial practical importance. The algorithms that suppress these harmful effects are called speech dereverberation algorithms [1], [4].

1.2. Direct Sound and Reverberation Components

Fig. 1.2 illustrates the reverberation produced by reflections of the wavefronts, which propagate outward from the source. The wavefronts reflect off the walls and superimpose at the microphone. In Fig. 1.2, this is illustrated by an example of a direct path and three reflections. Each of these wavefronts arrives at the microphone with different amplitude and phase. This is due to the fact that the length of the propagation paths to the microphone and the amount of energy absorbed by the walls are different. Therefore, as the term reverberation implies, in addition to the direct-path signal, the received signal contains delayed and attenuated copies of the source signal. More specifically, the received signal generally is described to be consisting of a direct sound, reflections that arrive shortly after the direct sound (commonly called *early reverberation/reflections*), and reflections that arrive after the early reverberation (commonly called *late reverberation/reflections*). The different sound components will now be discussed in more detail.

- *Direct Sound* is the first sound that is received at the microphone by passing the

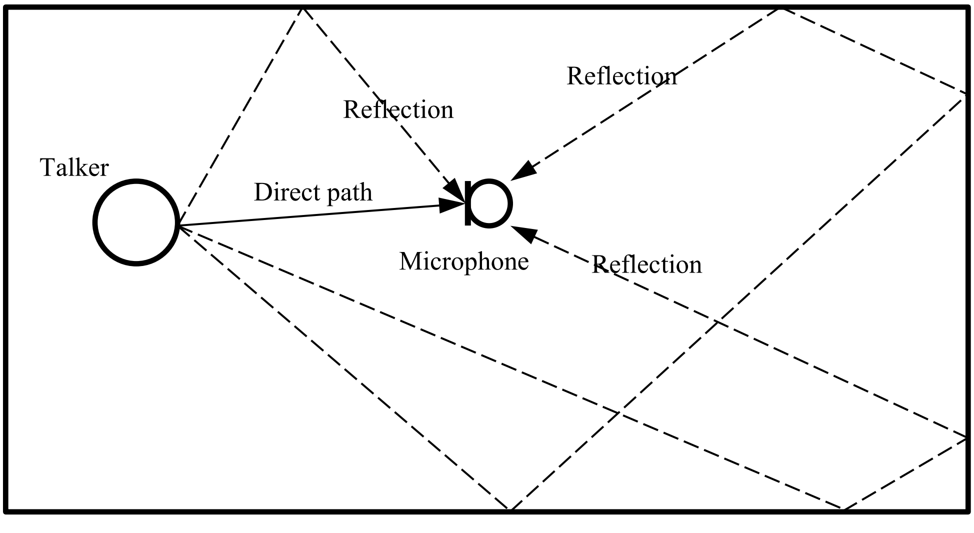


Fig. 1.2. Room reverberation illustration, including direct path and reflections [1].

direct path between the source and the microphone without reflection. The delay between the initial excitation of the source and its observation as the direct sound depends on the distance and the velocity of the sound.

- *Early Reverberations* are part of the reflections that are received during a short time after the direct sound. These components arrive at the microphone at different times and in different directions as compared to the direct sound and are also weaker in amplitude. So long as the delay of the reflections does not exceed a limit of approximately 80-100 ms with respect to the arrival time of the direct sound, early reverberation is not perceived as a separate sound from the direct sound [4]. Early reverberation is actually perceived to reinforce the direct sound and is therefore considered useful with regard to speech intelligibility [4]. This reinforcement is what makes it easier to hold conversations in closed rooms compared with outdoors. Early reverberation is mainly important in so-called small-room acoustics, since the walls, the ceiling and the floor are really close. On the other hand, early reflections cause a spectral distortion in the received signal, which is referred to as *coloration*. This effect is due to the short-term correlations introduced to the signal by early reflections. As a result, most of the dereverberation algorithms consider suppressing both the early and late reverberations. Furthermore, it should be noted that dereverberation algorithms have been proposed considering different applications including automatic speech recognition, where early reflections are not considered useful [1], [4].

- *Late Reverberations* are reverberation components that result from reflections which arrive with larger delays after the arrival of the direct sound. They are perceived by humans either as separate echoes, or as reverberation, and they degrade speech intelligibility [1], [4].

It should be noted that there is no clear boundary to distinguish between early and late reverberations and the definitions given above are highly comparative and relative. A typical notion is to consider this boundary at 50 ms after the direct path component.

The acoustic channel affecting the transition of the sound wave between a source and a microphone can be described by an impulse response known as the *acoustic impulse response* (AIR) or *room impulse response* (RIR). This impulse response represents the signal that is measured at the microphone in response to a source that produces a ‘sound impulse’.

Fig. 1.3 shows the simulated RIR for a room. As shown in the figure, the RIR is commonly split into three parts, the *direct path*, *early reflections*, and *late reflections*. The direct sound, early reverberations and late reverberations are, respectively, the product of the convolution of the three segments of the RIR with the clean signal. As can be seen from the figure, the energy of the reflections is reduced at an exponential rate. The notion of *reverberation tim*e has been developed based on this characteristic of the RIR. The reverberation time quantifies the severity of reverberation within a room, and is denoted by T60 or alternatively called RT60. Reverberation time is the time it takes for a 60 dB decay of the sound energy after switching off a sound source. The reverberation time is discussed in more detail in Chapter 2, Section 2.4.

When the distance between the source and the microphone varies, the proportion of

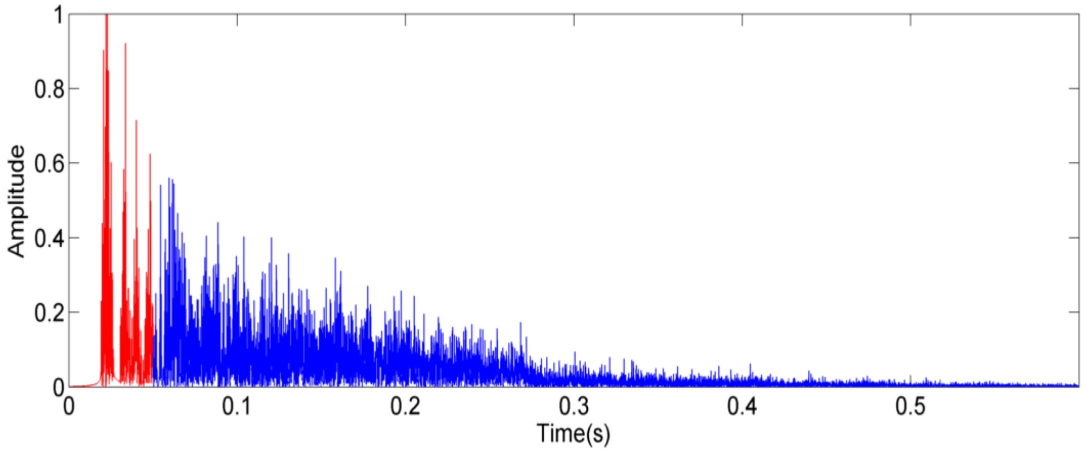


Fig. 1.3. Room impulse response for a room with reverberation time of 0.9 s. Red impulses are early reflections and blue impulses late reflection part. The strongest impulse is the direct path component [4].

the energy of the direct sound to that of the reflections varies accordingly. In other words, the energy of the direct sound changes with the distance between the microphone and the source, whereas the combined energy of the early and late reflections is approximately constant. The distance at which the direct path energy is equal to the ensemble energy of the early and late reflections is called the *critical distance* [4]. This means when the distance between a source and a microphone is greater than the critical distance, the overall energy of reflections is greater than the direct path energy. For further discussion and formulation of critical distance, the reader may refer to [4].

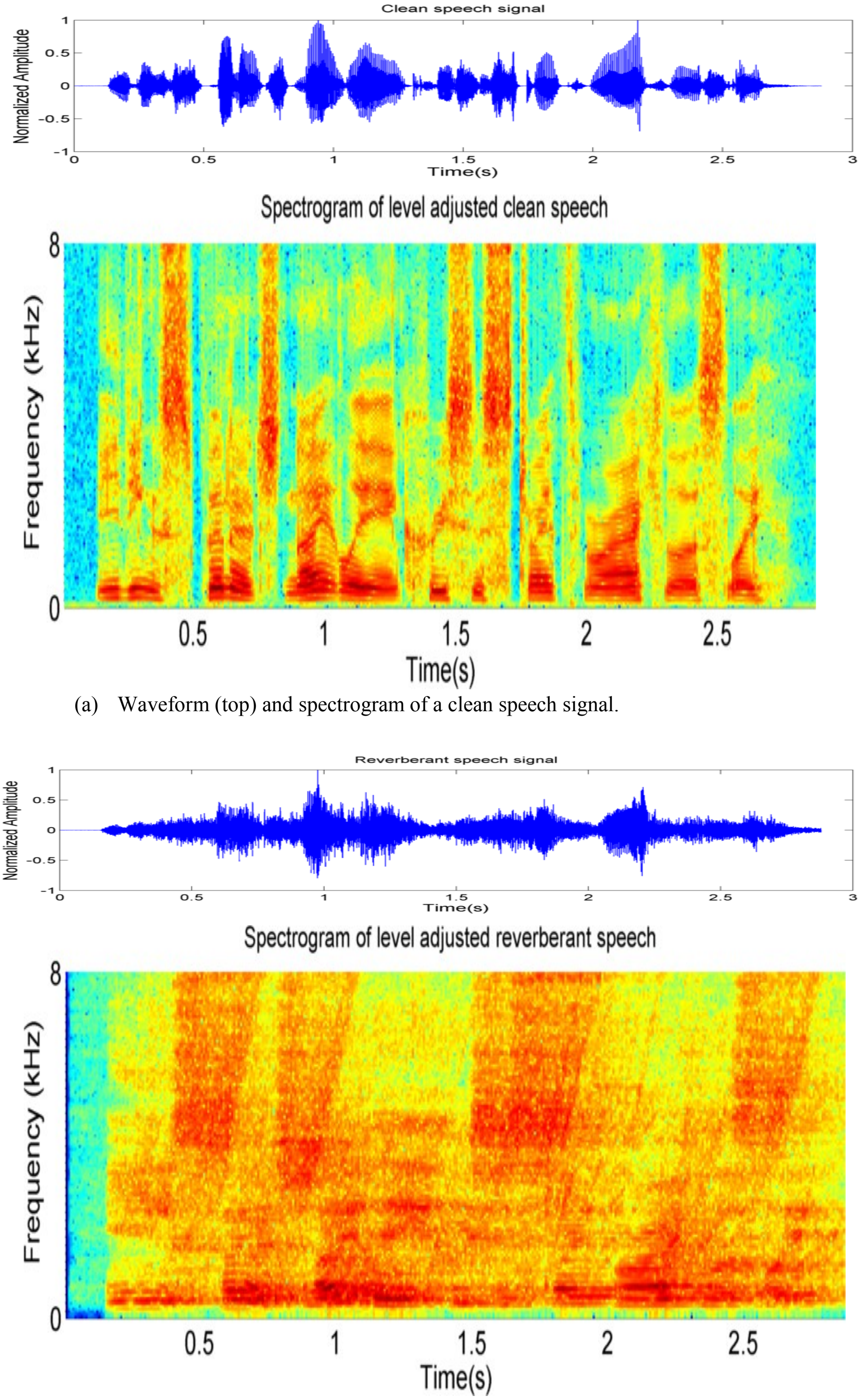
For development of effective dereverberation algorithms, it is of great importance to have a good understanding of the effects of reverberation on speech perception. This is discussed in the following section.

1.3. Effects of Reverberation on Speech Perception

The effects of reverberation on speech are illustrated in Fig. 1.4 through a clean speech utterance and the associated reverberant signal along with their spectrograms. The speech utterance is taken from the TIMIT speech database [5]. The *speech formants*, which are defined as the resonance frequencies associated with the vocal tract [6], are clearly detectable in the spectrogram of the clean signal. It is also visible that, in the anechoic signal, the speech phonemes are well distinguishable in time. To obtain the reverberant signal in Fig. 1.4 (b), the anechoic signal of Fig. 1.4 (a) was convoluted with a simulated room with reverberation time of 0.9 s. In the spectrogram of the reverberant signal, it can be clearly seen that the speech formants are blurred compared to that of anechoic signal. As well, both the spectrogram and the waveform show the smearing of the phonemes in time. Smearing of the phonemes causes the empty spaces between words and syllabi to be filled by reverberation which results in the overlap of subsequent phonemes. These distortions result in a degradation of speech intelligibility that is clearly audible. For a more detailed discussion on how dereverberation reduces the speech intelligibility, the reader is referred to [4].

1.4. Effects of Reverberation on Automatic Speech Recognition

One of the determining factors in the performance of automatic speech recognition (ASR) systems is the quality of the input speech signal. The performance of ASR systems tends to decrease rapidly when the distance between the source and the microphone increases. Consequently, when this distance increases, the signal to reverberation ratio (SRR) and the direct to reverberation ratio (DRR) decrease. The author in [4], by conducting an experiment on a simulated ASR system, has demonstrated that the word error rate (WER) of an ASR system increases rapidly for reverberation times larger than 0.2 s, and that the effects of reverberation on an ASR system are rather severe.



(a) Waveform (top) and spectrogram of a clean speech signal.

(b) Waveform (top) and spectrogram of the same speech signal when reverberated.

Fig. 1.4. A clean speech utterance from the TIMIT database and the associated reverberant speech signal along with their level-normalized spectrograms. The reverberant speech is produced by RIR with reverberation time of 0.9 s.

A block diagram describing an application of acoustic signal processing for cancelling the degradation effects on the speech signal is illustrated in Fig. 1.5. The source signal is the sound produced by the source, which is also the desired signal or the *anechoic or clean signal*. In addition to being 'transmitted' and affected by the *acoustic channel(s)*, the source signal is combined with the interfering signal(s) to be received as the microphone signal(s). The thick lines in Fig. 1.5 represent one or more signals, whereas the thin lines signify one signal. The interfering signals can either be interfering sounds or electrical interferences, such as sensor noise. The goal of the acoustic signal processor is to recover the desired signal by using the observed microphone signal. In this figure, reverberation is included as the effect of the channels on the source signal. In other words, in the specific case that noise and other interferences and various types of channel distortion are absent, the acoustic signal processor will be responsible only for the dereverberation task. As a result, this diagram can be considered as a general diagram for dereverberation as well.

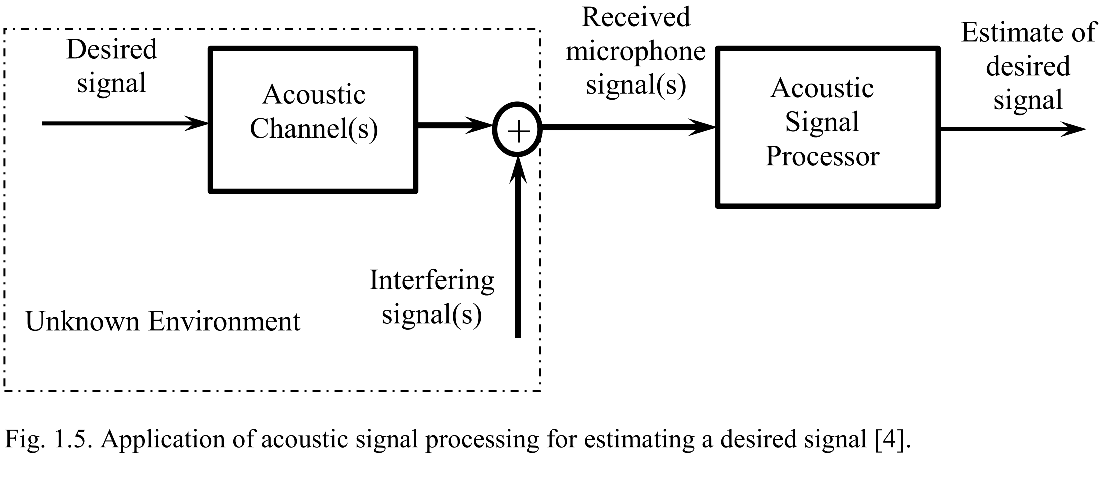

Desired signal
Acoustic Channel(s)
+
Received microphone signal(s)
Acoustic Signal Processor
Estimate of desired signal
Interfering signal(s)
Unknown Environment

Fig. 1.5. Application of acoustic signal processing for estimating a desired signal [4].

1.5. Motivation

One-microphone speech dereverberation, which is alternatively referred to as single-channel speech dereverberation, is the task of recovering the original anechoic signal (equivalent to the desired signal in Fig. 1.5) when only one observation of the reverberant speech signal (one microphone) is available. Clearly, in the dereverberation problem, as depicted in Fig. 1.5, the acoustic channel is unknown. Nevertheless, some methods take advantage of very limited knowledge about the channel. In the methods proposed in this work, however, no knowledge of the acoustic channel is used.

It is notable that single-channel speech dereverberation, in general, is considered a more difficult problem than multi-channel case since it does not allow for spatial processing across different observations of the signal [1], [4]. One should also note that, due to the same reason, multi-microphone algorithms are not usually applicable to single-microphone scenario; hence, the single-microphone case has to be separately addressed.

A number of important methods on single-channel speech dereverberation have been developed since about two decades ago. As one of the earliest major works on single-channel reverberant speech enhancement, in 1991, Bees *et al.* [7] proposed an algorithm which first estimated the cepstrum of the acoustic channel and then used a least squares technique for inversion. Although their results of channel-estimation are satisfactory, they are derived for minimum-phase responses or for mixed-phase responses having a few zeros outside the unit circle, which are not realistic. Authors in [4], [8], and [9] have developed dereverberation algorithms based on the effects of reverberation on modulation transfer function (MTF). However, this method has limited applicability since it is based upon the assumptions that do not necessarily match the features of real speech and reverberation. Firstly, real speech signals were not considered. Secondly, a simple exponential model was employed for modeling the RIR. In [10] and [11], the authors employ the harmonic structure of speech for dereverberation. By using this method, good results are achieved, but the algorithm involves producing a large amount of reverberated speech using a fixed RIR.

By assuming that late reverberation components are independent of early reverberation components, some researchers have focused only on the removal of late reverberations by using the so called spectral enhancement methods. This is done by using short-time Fourier transform (STFT) by estimating the short-term power spectral density (STPSD) of the late reverberation components so as to perform magnitude subtraction without any phase correction. The main challenge in such methods is the estimation of the STPSD of the late reverberant speech components from the observed reverberant signal. In this category of methods, several techniques have been proposed for the estimation of the STPSD of the late reverberations [4], [12], [13]–[17]. Spectral subtraction is a commonly used technique for dereverberation. In terms of computational complexity, it is relatively less complex; it can be used in real time applications, and results in the suppression of both the background noise and late reverberation. Nevertheless, the first drawback of this category of methods is that it simply does not consider the early reverberations while they are especially important for automatic speech recognition applications, which are sensitive to short reverberations. In addition, due to nonlinear filtering in these methods, artifacts such as musical noise[1] are introduced and these are typically annoying. Moreover, in these methods, a priori knowledge of the RIR (i.e., the reverberation time) is usually required, in which case these techniques resort to blind reverberation time estimation techniques to achieve a complete blind dereverberation.

Yegnanarayana and Murthy [18], [19] observed that the LP residual of reverberated speech is smeared and resembles Gaussian noise, while that of clean voiced-speech shows patterns of damped sinusoids within each glottal cycle. Based on this result, they estimate the LP- residual of clean speech and then synthesize an enhanced speech. Their method identifies and manipulates the LP-residual based upon the regions of reverberant speech with different SRR, namely, high SRR, low SRR and pure reverberation. As a result, this is a temporal domain method which mainly enhances the speech in the high SRR regions. Authors in [20], combined a similar LP-residual based approach to enhance reverberant speech in the high SRR regions, with spectral subtraction to reduce late reverberation.

Gillespie *et al*. [21] made an important observation that the kurtosis of LP residuals could be a reasonable measure of reverberation. They used kurtosis maximization of LP residual of the reverberant signal as a criterion for adjusting the weights in their inverse filter. This observation has been used in a number of algorithms proposed later (e.g. [22] and [23]). This inverse-filtering method, however, is merely effective for suppressing the short reverberation component.

[1] In the spectral subtraction methods, musical noise is caused by spurious peaks introduced to the spectrum of the speech signal due to errors in noise or SNR estimation. When the enhanced signal is reconstructed in time domain, these peaks result in short sinusoidals whose frequencies vary from frame to frame. This produces a noise which is audible particularly in low SNR regions and silent gaps where it is not masked by the speech signal [1].

Most single-microphone dereverberation methods developed so far have aimed at reducing effects due mostly to late reverberations. This is while the frequency response of early reverberations is rarely flat, meaning that it distorts the speech spectrum and reduces speech quality, particularly for ASR applications [24].

As joint dereverberation of both early and late components is quite challenging, very few single-microphone two-stage algorithms have appeared in the literature to this goal. Wu and Wang [22] used the method by Gillespie *et al.* [21] as the first stage of their algorithm, and followed it by spectral subtraction to reduce late reverberation. However, their method yields satisfactory results only when the reverberation time is short (i.e. less than 0.4 s). Also, noisy environment has not been considered in their work. In a similar approach in [25], temporal averaging to suppress early reflections was combined with spectral subtraction.

In a very recent paper [26], the authors have employed skewness maximization of the LP-residuals of the reverberant signal, rather than the kurtosis maximization, as a criterion for adjusting the weights in the inverse filter. They pointed out the reason for such a preference as follows: in high reverberation times, the kurtosis-based objective function for adaptive inverse filtering has many saddle points (along with the maximum points), and convergence is usually to one of them, leading to an inaccurate filter estimate. However, for speech dereverberation applications, their algorithm is not very effective, especially for long reverberations, as it is based on a single-step LP-residual inverse filtering, which cannot suppress both long and short reverberations at the same time. Kinoshita *et al.* [27], on the other hand, proposed an algorithm consisting of LP-based spectral subtraction followed by a cepstral mean subtraction (CMS). Their algorithm is fast, but fails to sufficiently estimate the late reverberation spectra in single-channel implementation. As a result, it is not sufficiently effective in the single microphone case.

1.6. Objective of the Thesis

The objective of this thesis is to develop new algorithms to improve the efficiency of single-channel dereverberation. The algorithms proposed in this thesis are based on a two-stage development of inverse-filtering by using LP-residuals followed by spectral enhancement. The proposed algorithms are designed so that the long reflections are also suppressed in the first stage, i.e., inverse-filtering. This is done by using a linear prediction scheme which includes prewhitening followed by a delayed long-term linear prediction. The difference in the two proposed algorithms is that one uses kurtosis maximization, whereas the other utilizes skewness maximization in order to control the weight updates of the inverse filter. Clearly, because of the difference in the behaviour of the kurtosis and skewness of the LP-residuals of reverberant signals, some parameters are also different in implementing the two algorithms. The second-stage of the proposed algorithms is identical to that of Wu and Wang [22]. However, the resulting two-stage algorithms are more effective in suppressing the long reflections, which are the main source of degradation of speech signal, while keeping the efficiency for short reflections.

1.7. Thesis Organization

This thesis is organized as follows. In Chapter 2, theoretical background about speech dereverberation is first given. This begins with the description of a system representation for the general problem of reverberation. Then the concept of AIR or RIR and its different parts are introduced and explained. Reverberation time, as a measure for the severity of reverberation in an RIR is then described. Next, statistical modelling of reverberation is introduced in order for the reader to have more insight to reverberation. The next section of this chapter is devoted to dereverberation evaluation. Some of the qualitative, subjective and objective measures of reverberation are explained in this section. These measures are the ones that have been used in, or are related to, the evaluation of the proposed algorithms in Chapter 4. They have been chosen based upon the nature of the proposed algorithms to be comparable to similar works in the literature. In the next section, an overall classification of the dereverberation algorithms is given; this classification is based on the level of the channel and source knowledge and the difference in the signal processing techniques utilized. Finally, a review of the most relevant dereverberation methods is given.

Chapter 3 describes the two new algorithms developed in this work. This chapter starts with the introduction which is a review to the previous works related to the algorithms proposed in this thesis. Then the formulation of the single-channel dereverberation in the proposed algorithms is described. The next subsections are devoted to describing the different parts of the algorithms which are the multiple-step linear prediction, the inverse-filtering by maximization of kurtosis and skewness and the spectral subtraction.

Chapter 4 is concerned with the performance evaluation of the proposed algorithms and comparison with the existing works. In this chapter, the experimental setup and the parameters used in implementing all the algorithms are explained first. The results of the algorithms to different quantitative and qualitative measures are then one by one described and compared to two existing major single-channel dereverberation algorithms, which are among the most successful and most cited ones for single-channel speech dereverberation. The algorithms are compared in terms of their equalized impulse responses and their energy decay curves, normalized segmental SRRs, ASR test, perceptual evaluation of speech quality (PESQ) and spectrograms. The robustness of the proposed algorithms against background noise is also compared to the reference algorithms.

In Chapter 5, the thesis is concluded by summarizing the results obtained and discussing the possibilities for further future work.

Chapter 2

Theoretical Background and Literature Review

2.1. Introduction

This chapter aims to briefly introduce some of the main aspects of the reverberation and dereverberation that are directly linked to the study of the algorithms proposed in this thesis in Chapter 3. Towards this goal, the general problem formulation of reverberation is first introduced. Then, the concept of AIR and its pertaining characteristics are explained. Next, the concept of reverberation time, and the relevant theory and measurement are briefly explained. Afterwards, in order to grasp more insight into the reverberation phenomenon, in contrast to the typical time domain modeling, a statistical modeling of reverberation is also briefly presented. Following this theoretical background, some of the various ways of evaluating dereverberation are briefly explained. This includes only those measures that are used in, or directly connected to, the evaluation of the algorithms in this thesis in Chapter 4. The most relevant measures have been chosen based upon the nature of the proposed algorithms and similar works in the literature. Finally, a broad classification of dereverberation algorithms is given followed by a brief introduction and explanation of some of the major dereverberation algorithms that are most relevant to the methods proposed in this thesis.

2.2. System Description

Figure 2.1 illustrates a generic system diagram for multichannel dereverberation. The single-channel scenario would be when there is only one acoustic channel and one microphone. The speech signal, $s(n)$, propagates through acoustic channels, $H_m(z)$ for m = 1 to M, and is collected at the output by using M microphones to result in signals $x_m(n)$. The noise in the system is assumed additive and is represented by $v_m(n)$.

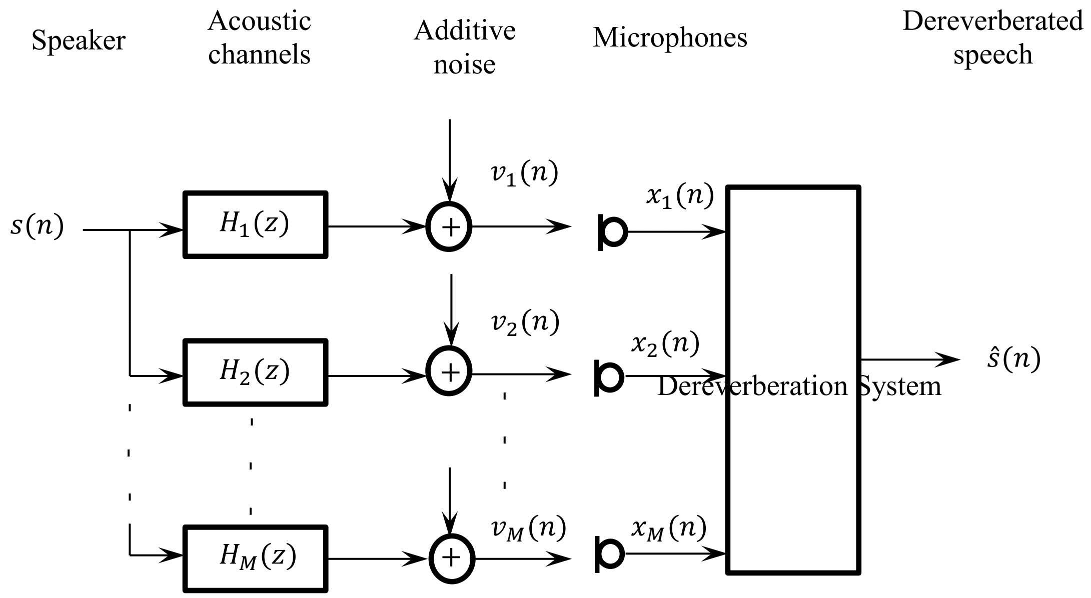


Fig. 2.1. General multi-channel reverberation-dereverberation system model [1].

The observed signal, $x_m(n)$, at microphone m is the superposition of

(i) The direct-path signal, which travels the direct path from the talker to the microphone arriving with attenuation and propagation delay

(ii) A theoretically infinite set of reflections of the original signal arriving at the microphones at later time instances whose attenuation is dependent on the properties of the reflecting surfaces. This can be expressed as

$$x_m(n) = \sum_{i=0}^{\infty} h_{m,i}(n)s(n-i) \tag{2.1}$$

where $h_m(n)$ is the impulse response of the acoustic channel from the talker to the *m*-th microphone. In other words, $h_{m,i}(n)$ represents the attenuation and the propagation delay corresponding to the direct signal and all the reflected components for the signal observed at the *m*-th microphone [1], [28].

The aim of speech dereverberation is to find a system that by observing $x_m(n),\ m = 1, \dots, M$ as the input, obtains the output $\hat{s}(n)$ which is a 'good' estimate of $s(n)$. How and when $\hat{s}(n)$ is considered a 'good' estimate of $s(n)$, depends on the application. For instance, it may be desired to estimate $s(n)$ by using minimum mean square error (MMSE) criterion. However, for speech dereverberation, other criteria may be more relevant, such as those related to perceptual quality [1], [29]. Speech dereverberation is a blind problem since the goal is to recover the original signal $s(n)$ when the acoustic channels, $H_m(z)$'s, are unknown.

Recently, efforts in acoustic signal processing have led to several algorithms for speech dereverberation and reverberant speech enhancement. Consistent with [1], in a broad sense, all speech dereverberation methods fit into one of the three main categories described below:

1. *Beamforming* – In this approach, an array of microphones is used and the observed reverberant signals arrive at the different microphones with different delays and attenuations. The array of microphones might have different shapes such as a line array or a circular or a 3-D shaped array. The received signals are filtered and

weighted so as to form a beam of enhanced sensitivity in the direction of the desired source (so called direction of arrival, DOA) and to attenuate sounds from the other directions. Clearly, beamforming is dependent on the availability of multi-microphone inputs. Beamforming is a multiple-input single-output process.

2. *Speech enhancement* – In these methods, according to an *a priori* defined model of the speech signal or spectrum and using some features of the clean speech signal as compared to the reverberant signal, the speech signals are enhanced. Although many speech enhancement techniques benefit from the use of multiple inputs, speech enhancement is often a single-input single-output approach [1].

3. *Blind deconvolution* – An inverse filter is estimated blindly to compensate for the effect of the acoustic impulse response on the speech signal and recover the original signal. In some cases the acoustic impulse responses are identified blindly and then the inverse filter is built, whereas in other cases the inverse filter is not shaped by estimating the acoustic impulse responses, but by using some other features such as those of the LP-residual signals.

2.3. Acoustic Impulse Response

The acoustic impulse response (AIR) is the impulse response that describes the acoustics of a given enclosed space which in case of a room is called room impulse response (RIR). Consequently, a natural approach to dereverberation is to estimate the AIR (RIR) that has affected the signal. For that purpose, and also to have a good viewpoint of reverberation and dereverberation, it is necessary to study some characteristics of the AIR. Herein, the focus is on the RIRs, where reverberation has a substantial effect on telecommunication applications.

The room impulse response has been modelled in several different ways including both finite impulse response (FIR) and infinite impulse response (IIR) structures. The choice of the RIR model will generally influence the algorithmic development. One way of describing RIR is to use the definition of reverberation time, which was originally introduced by Sabine [30]. The reverberation time, $RT60$, is defined as the time taken for the reverberant energy to decay by 60 dB once the sound source has been abruptly shut off [1]. The geometry of the room and the reflectivity of the reflecting surfaces are the factors that determine the reverberation time of a room. When measured at a fixed location in a room, the reverberation time and the RIR are approximately constant. However, they vary as the talker, the microphones or other objects in the room change location [31]. In particular, as the talker-microphone distance increases, the proportion of the energy of the direct-path component to that of reflection components of RIR varies. The distance at which these two energies become equal is called the *critical distance* [1].

Figure 2.2 shows an example of the room impulse response extracted from MARDY database [32]. First, there is an initial dead time related to the time it takes for the sound to travel the direct path between the source and the microphone. This short period of near-zero amplitude, which is sometimes referred to as the direct-path propagation delay, is followed by a peak. Depending on the source-microphone distance and the reflectivity of the surfaces in the room, the amplitude of this peak due to direct-path propagation may be greater or less than the amplitude of the later reflections. The example of Fig. 2.2 shows a RIR with a strong direct-path component. This indicates that the source-microphone distance is relatively short.

The early and the late reflections are separated in the figure with two different colors. The early reflections are often taken as the first 50 ms of the impulse response [31], and consist of impulses of relatively large magnitude compared to the late reflections. The propagation of the wave from the speaker's lips to the microphone can be represented by the convolution of the speech signal with the RIR. The RIR early reflections cause spectral changes in the sound resulting to a perceptual effect that is called coloration [1], [31]. In general, it has been shown that early reflections can have a positive impact on the intelligibility of the speech in a way similar to reinforcing the direct-path component [1], [31], [33]. This is due to the characteristics of the human hearing system in which the closely spaced echoes are not distinguished due to the masking properties of the ear. However, coloration can degrade the quality of recorded speech [31]. Hence, the dereverberation algorithms have to take care of both short and long reflections, especially when non-human hearing is of importance such as in automatic speech recognition systems.

The late reflections, which are also referred to as the tail of the impulse response, are the closely spaced, decaying impulses that follow the early reflections. The late

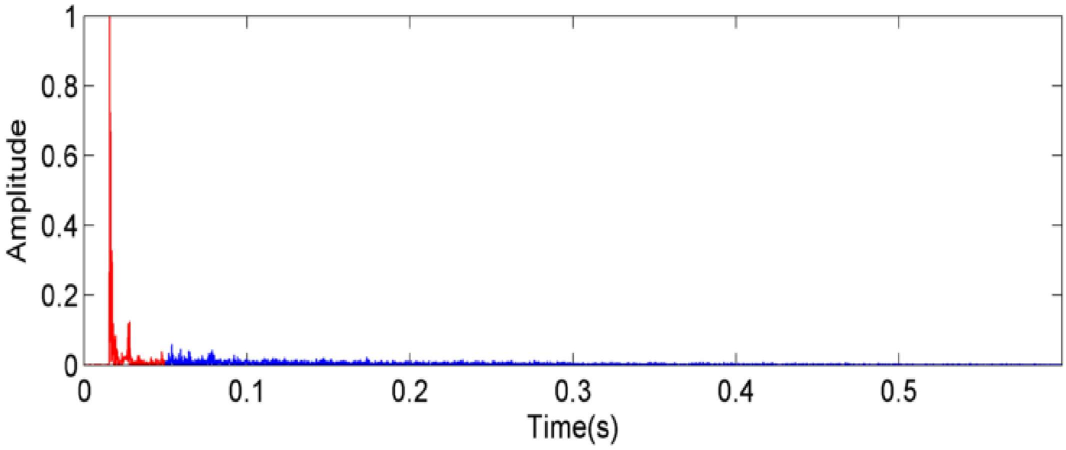


Fig. 2.2. An example room impulse response for a room extracted from MARDY database [32].

reflections produce effects of a ‘distant’ and ‘echo-ey’ sound and provide the major contribution to what is generally understood as reverberation in everyday experience. They are the main source of degradation in the quality of speech sound although, depending on the application, the early reflections are also, at least partially, considered harmful [1], [4], [31].

In terms of spectral characteristics, the effect of the room can be represented as the room transfer function. The properties of the room transfer function have been studied extensively in the room acoustics literature. As an important property, Neely and Allen [34] concluded that the RIRs in most real rooms possess non-minimum phase characteristics.

Room transfer functions are generally stable with the impulse response coefficients $h_{m,i}(n)$ tending to zero with increasing index i. Therefore, it is sufficient to consider only the first L_h coefficients in (2.1) [1]. The choice of L_h is often related to the reverberation time of the room. Taking into account any additive noise sources, the observed signal at the m-th microphone can be written in a vector form as

$$x_m(n) = h_m^T(n)s(n) + v_m(n) \tag{2.2}$$

where $h_m(n) = [h_{m,0}(n)\ h_{m,1}(n) \dots\ h_{m,L_h-1}(n)]^T$ is the L_h-tap impulse response of the acoustic channel from the source to microphone m, $s(n) = [s(n)\ s(n-1) \dots\ s(n-L_h+1)]^T$ is the speech signal vector and $v(n)$ is the observation noise. Equation (2.2) also corresponds to Fig. 2.1, where, interference is also taken into account in the reverberation scheme.

2.4. Reverberation Time

As mentioned earlier, the reverberation time is a parameter defined for describing the reflectivity of an acoustic enclosed space. To measure the reverberation time of a room, first the room is excited by a broadband signal until a steady-state uniform sound energy distribution is achieved. Then, the sound source is abruptly switched off and the resulting decay of squared sound pressure is recorded. By plotting this energy decay versus time, a curve is obtained which is known as the energy decay curve (EDC). The reverberation time, $RT60$, is defined as the time in seconds required for the EDC to decay by 60 dB [1].

The definition of reverberation time originates from the early work of Sabine [35] who concluded that the reverberation time was proportional to the volume of the room, and inversely proportional to the amount of absorption in the room [1]. Based on his method, by neglecting the effect of attenuation due to propagation through the air, the reverberation time is estimated as

$$RT60 = \frac{24\ln(10)}{c} \frac{V}{\alpha_{Sabine} A} \quad s. \tag{2.3}$$

where α_{Sabine} represents the total absorption in the room calculated by summing the products of Sabine's absorption coefficients and their corresponding areas (for more information see [1], [35]).

The reverberation time is alternatively given by Eyring's reverberation formula [35] as

$$RT60 = -\frac{24\ln(10)}{c} \frac{V}{\ln(1-\alpha_{Eyring})\,A} \quad s, \tag{2.4}$$

where α_{Eyring} is the Eyring sound absorption coefficient similar to that in the Sabine's method.

Both the Sabine and the Eyring reverberation times may also be calculated using an average absorption coefficient and a total corresponding reflecting surface area. Furthermore, the Eyring absorption coefficients can be derived from the Sabine coefficients [1].

When the average absorption coefficient, $\bar{\alpha}$, is small, by using the expansion

$$-\ln(1-\bar{\alpha}) = \bar{\alpha} + \frac{\bar{\alpha}^2}{2} + \frac{\bar{\alpha}^3}{3} + \cdots \tag{2.5}$$

it can be shown that Eyring's and Sabine's reverberation times become approximately equal. In addition, these expressions indicate that the reverberation time of the room is independent of the locations of the source and the microphones [1].

If the RIR is known, by definition, the EDC can be obtained from the Schroeder integral [35]

$$EDC(t) = \int_t^{\infty} h^2(\tau) d(\tau) \tag{2.6}$$

where $h(t)$ is the impulse response of the room. The integral in (2.6), calculates the sum of the energies of the impulses after time t.

An example is given in Fig. 2.3, which shows the EDC for a measured impulse response. The reverberation time $RT60$ can be obtained by using an EDC plot only if the impulse response is measured at a distance greater than the critical distance. This is because $RT60$ is independent of any effects of the direct path component such as the geometry of the source and the microphones which are present at shorter distanc1es. In addition, for the estimation of $RT60$, measurements should be performed at levels greater than the ambient noise level in order to avoid the effects of such noise. Considering these factors, useful estimates of $RT60$ can be obtained from EDC plots such as Fig. 2.3 by measuring the slope of only the free decay section, this being the part that has a near constant gradient. In Fig. 2.3, the estimated reverberation time by this method, so called the Schroeder method, is 0.52 s.

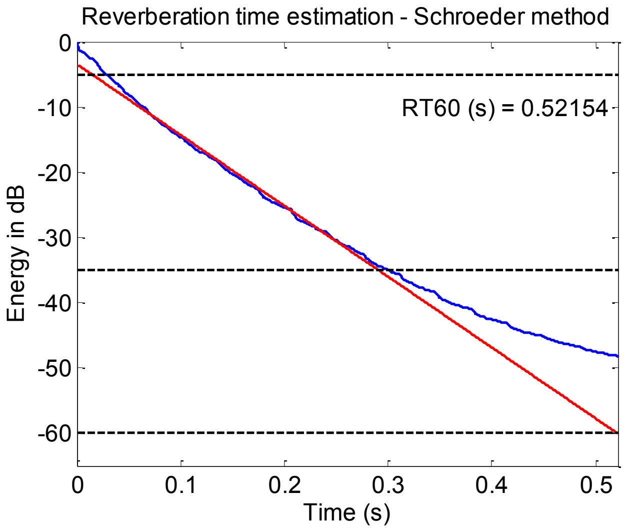


Fig. 2.3. The EDC curve and the tangent line for RT60 calculation.

2.5. Statistical Modeling of Reverberation

Time domain modelling of reverberation described by (2.1) or (2.2) is the first type of description that intuitively strikes one’s mind. However, in addition to this fundamental description, reverberation has been also modelled by using some statistical approaches that have proved to be useful.

First, Moorer [36] suggested that the reverberation effect can be produced by the convolution of a clean speech with a Gaussian noise modulated by exponentially decaying envelope. Polack [37] then proposed modeling the RIR as the product of a stationary Gaussian noise process and an exponentially decaying envelope:

$$h(t) = b(t)e^{-\Delta t}, \quad t \geq 0 \tag{2.7}$$

where $b(t)$ is a zero-mean Gaussian stationary noise, and Δ is the exponentially decaying parameter which is related to the reverberation time, $RT60$, by

$$\Delta \ = \frac{3ln(10)}{RT60} \tag{2.8}$$

Since reverberation time is frequency dependent, the model described by (2.7) can also be implemented in separate acoustic frequency bins as

$$h_k(t) = b_k(t)e^{-\Delta_k t}, \qquad t \geq 0\,, \qquad k = 1,2, \dots, K \tag{2.9}$$

This model works well when the distance between the source and the microphone is larger than the critical distance. For shorter source-microphone distances, Habets [12] proposed a more accurate model as:

$$h_k(t) = \begin{cases} b_d(t), & 0 \leq t < T_r \\ b_r(t)e^{-\Delta t}, & T_r \leq t < \infty \\ 0, & otherwise \end{cases}$$

where $b_d(t)$ and $b_r(t)$ are two zero-mean mutually independent and identically distributed (i.i.d.) Gaussian random variables, and T_r is the time (with respect to the arrival time of the direct sound) at which it is assumed that the late reverberation starts.

2.6. Evaluation of Dereverberation

Speech dereverberation is only one of the domains where signal processing helps enhance the quality of speech signals. Speech quality measurement, in general, is performed either by subjective or objective evaluation. However, evaluation of speech dereverberation is a more specific case. Subjective and objective measures of speech quality and speech dereverberation will be briefly discussed in this chapter.

Objective quality measures are typically classified into intrusive and non-intrusive measures. In intrusive measurement, the processed (or distorted) signal is compared to an undistorted (reference) signal. In speech dereverberation, this means comparing the processed signal by the algorithm with the clean signal which has no reverberation. In contrast, in non-intrusive measurement, the evaluation is performed by using merely the distorted (processed) speech. Typically, non-intrusive quality measures are only used when access to the reference signal is impossible. This is because not having access to the reference signal makes the evaluation more complex. Thus, in this section and throughout this work, the assumption is that the reference signal is available meaning that the measures are intrusive.

Speech quality measurement, on the other hand, can be classified into qualitative and quantitative evaluation. Qualitative evaluations include quality measures that use visualization of the resulting signals or impulse responses such as spectrograms, and equalized room impulse responses, while quantitative measures are those that perform the assessment by assigning a score to the signal under evaluation.

Owing to the fact that the degree of correlation of different general speech quality measures with speech reverberation, as a specific case, is different, reliable quantitative measurement of reverberation level of speech signal is still difficult, and a solid universally-accepted methodology has not yet emerged. In other words, an objective measure is considered highly reliable for dereverberation only if it shows high correlation with subjective tests. Developing quality measures for dereverberation, which are more and more correlated with subjective assessment is a subject of research (see [38] for example). Nonetheless, existing objective measures are usually combined together to evaluate the performance of speech dereverberation algorithms.

2.6.1. Qualitative Evaluation by Visual Representation

Speech Waveform and Spectrogram

The speech waveform and the spectrogram are often used for representing the speech signals visually and comparing them with each other. Spectrogram is the time-frequency visualization of the power spectral density (PSD) of the signal in which one axis (usually horizontal) is assigned to time and the other axis represents the frequency. In other words, it illustrates the alterations of the power of the speech signal in different frequencies through time by using a color-map scheme in which different colors indicate different energy levels.

The smearing effect of reverberation is clear in the waveform and in the spectrogram of speech. However, it is usually difficult to detect how severely the signal is degraded in a relative sense, especially when the reverberation levels of the two signals are not so apart.

Equalized RIRs

For the inverse-filtering algorithms, one of the other visual evaluations of the results is using the equalized RIRs. The equalized RIRs are obtained by convolving the derived inverse-filter into the original RIR. Plotting and comparing the shape of the equalized RIRs and considering how the impulses are suppressed in different parts is a qualitative evaluation for inverse-filtering. This will be used in Chapter 3 of this work.

2.6.2. Subjective Measures

Subjective speech quality measurement is performed by using human participants to rate the quality of speech signals by assigning scores to them in an opinion scale. The most commonly used subjective quality measures for speech transmission over voice communication systems have been standardized by the International Telecommunications Union (ITU-T). Subjective speech quality measures are twofold; conversational and listening-only tests. For both types, a 5-point opinion scale, from bad to excellent, is recommended to use, known as listening quality scale [39]. Another speech quality scale, used for listening-only tests, is the listening-effort scale. As a third measure, a binary opinion scale is usually employed for conversational tests. These scales are listed in Table 2.1 [4].

In a listening test, subjects listen to the recordings degraded by an acoustic channel, channel, and enhanced by the algorithm under test. Then, depending on the type of the test, the subjects grade the quality of each signal or the effort required to understand it. In conversational tests, subjects are asked to use a voice communication system through a conversation and provide their opinion on its quality. The average opinion score across all the subjects is then calculated which is known as mean opinion score (MOS). This score represents the subjective quality of the algorithm under evaluation. The more the number of subjects used for testing, the more realistic the opinion score

Table 2.1. Subjective speech quality measurement scales recommended by ITU-T [39].

Listening-Quality Scale:

Quality of the speech/connection	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Listening-Effort Scale:

Effort required to understand the meaning of sentences	Score
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

Conversation Difficulty Scale:

Did you and your partner have any difficulty in hearing over the connection?	Yes 1 / No 0

becomes. This makes it cumbersome and time-consuming to perform such an evaluation. Furthermore, even by using a large number of subjects, the MOS variance can still be high, which is another disadvantage of this type of assessment. In addition, the expected quality of the speech signals can be different depending on the application. For instance, the expected speech quality for a cheap ordinary mobile telephone device would be much lower than that of a modern expensive conference system. Due to the constraints mentioned above, it would be more practical if an automatic speech evaluation system would exist by which the quality measures could be obtained [4].

2.6.3. Objective Measures

Based upon the preceding subsection and with the ever-evolving voice communication systems nowadays, an increasing demand for robust objective speech quality measures that correlate well with subjective tests is felt. Objective quality measures are helpful evaluation tools during the design and validation of algorithms, codecs and communication systems. Based on different speech analysis models, various objective measures have been developed by researchers over the last two decades [4].

During the design and validation stages of algorithms, codecs, and communication systems, objective quality measures are valuable assessment tools. Over the last two decades, researchers have developed different measures based on various speech analysis models [40], [41].

Objective speech quality measures, in general, are typically classified into three domains: time domain, spectral domain or perceptual domain. The time domain measures are generally applicable to analogue or waveform coding systems, where the receiver reproduces the waveform. Nevertheless, they can also be used to determine the improvement in the speech quality. Signal to noise ratio (SNR) and segmental SNR are typical time domain measures [4], [42]. Since the spectral domain measures are less influenced by the possible misalignments between the original and the processed signal, they are usually preferred to time-domain measures. Perceptual domain measures, which are developed based on models of the human auditory system, are known to have a higher chance of predicting the subjective quality of speech compared to time and spectral domain measures. Theoretically, perceptually relevant information is both sufficient and necessary for a precise evaluation of perceived speech quality [4], [40].

Considering the facts mentioned above, it is not surprising that most of the objective measures are intrusive and perceptually based. These measures usually follow psychoacoustic considerations and are trained on subjective databases to become as close as possible to human perception. One of the perceptual measures of speech quality is the one that ITU-T has standardized as perceptual evaluation of speech quality (PESQ) in 2001 as ITU-T Recommendation P.862 [4], [43]. PESQ was originally developed to evaluate the listening quality of a speech signal degraded by codecs, background noise and packet loss.

As mentioned earlier, among the objective measures, intrusive measures are those that use the comparison of the processed signal to a reference signal. Intrusive measures can be classified into three categories. The three categories include perceptually-based measures, channel-based measures, and measures that are based on neither of the two.

a) Intrusive Waveform-based Measures

One of the most important and most relevant speech quality measures for dereverberation evaluation is segmental signal to reverberation ratio [4]. This quality measure is used in this work and is introduced below.

Segmental Signal-to-Reverberation Ratio

Similar to segmental SNR [42], the instantaneous segmental signal to reverberation ratio (SRR) [44] of the m^{th} frame is defined as

$$SRR_{seg}(m) = 10\ \log_{10}\left(\frac{\sum_{n=mR}^{mR+N-1} z_d^2(n)}{\sum_{n=mR}^{mR+N-1}\left(z_d(n) - \hat{z}_d(n)\right)^2}\right) \quad [dB], \tag{2.10}$$

where N is the frame length, normally such that Nf_s is equal to 32 ms (this is the time interval in which the speech signal can be assumed to be wide sense stationary), R is the frame rate, m is the frame number, $z_d(n)$ is the delayed version of the anechoic (clean) signal, which is noted as the direct signal, and $\hat{z}_d(n)$ is the enhanced (processed) signal. The frame rate depends on the overlap between adjacent frames, which is usually chosen between 50 to 75 %. After calculating the SRR of frames, the final score, the mean segmental SRR, is then obtained by averaging the SRR scores over all the frames.

b) Intrusive Perceptually-based Measures

Bark Spectral Distortion

The Bark spectral distortion (BSD) is one of the extensively used speech quality measures that are based on the models of the human hearing system [45]. According to the studies, this measure has a very high correlation with MOS scores (subjective assessment) [45], [46]. The BSD is based on using the Bark spectra of the direct signal,z_d, and the enhanced signal, $\hat{z}_d$. These spectra are respectively denoted as $L_{z_d}$and $L_{\hat{z}_d}$. The BSD score is calculated using [4]

$$BSD = \frac{1}{m} \sum_{m=0}^{M-1} \frac{\sum_{k_s=1}^{K_s} (L_{z_d}(m, k_s) - L_{\hat{z}_d}(m, k_s))^2}{\sum_{k_s=1}^{K_s} (L_{z_d}(m, k_s))^2}, \tag{2.11}$$

where m and k_s denote the frame number and the Bark frequency bin, respectively.

The modified Bark spectral distortion (MBSD) further adds another step in calculating the Bark spectra by considering a noise-masking threshold [47]. The aim of this threshold is to differentiate between the audible and inaudible distortions. In this measure, it is assumed that the parts of the speech whose loudness falls below the noise masking threshold are inaudible and are thus neglected in the calculation of the perceptual distortion. As well, the MBSD makes use of a simple cognition model to calculate the distortion value [47].

In a more recent improvement to MBSD, the enhanced modified Bark spectral distortion (EMBSD) measure has been introduced [48]. This new measure develops a more complex cognition model for calculating the distortion value, which is based on removal of a couple of assumptions in MBSD that seem not to be met in some conditions. These conditions include a speech utterance containing background noise or a speech utterance with distortions such as bit errors or frame erasures encountered in real network environments. In EMBSD, for a better cognition model, a couple of psychoacoustic results have been extracted from the literature and incorporated into the cognition model (for further study see [48]).

Perceptual Evaluation of Speech Quality

As mentioned earlier, perceptual evaluation of speech quality (PESQ) is the objective measure recommended by ITU-T in P.862 (February 2001) [49]. The PESQ is a rather complex measure which is the result of several years of development and is applicable to speech codecs as well as intrusive measurements. The PESQ can be applied to real systems that include filtering and variable delay, as well as distortions due to channel errors and low bit-rate codecs. It is notable that, prior to the PESQ, the PSQM measure, which was recommended by ITU-T P.861 (February 1998), was only applicable to speech codecs without being able to take care of filtering, variable delay and short localized distortions into account. The PESQ, in contrast, accounts for these effects with transfer function equalization, time alignment, and a new algorithm for averaging distortions over time. In P.862, the PESQ score is recommended to be used for speech quality assessment of 3.1 kHz (narrow-band) handset telephony and narrow-band speech codecs.

PESQ compares an original signal $s(t)$ with a degraded signal $z(t)$, obtained by passing $s(t)$ through a communication system, or with the enhanced signal $\hat{s}(t)$ produced by the enhancement system. PESQ gives a prediction of the perceived quality that would be given to the signal by subjects in a subjective test.

PESQ first computes a series of delays between the original signal and the signal under test. These delays are calculated for each time interval whose delay is significantly different from the previous time interval. A start and stop point is assigned to each of these time intervals. This alignment algorithm works based on the principle of comparing the confidence of having two delays for a certain time interval with the confidence of having a single delay for that interval [4]. The algorithm follows delay changes both during the silent frames and during active speech frames. By using a perceptual model, based on the set of delays that are found, PESQ compares the original signal with the aligned signal under test. This process is based upon transformation of both the original and the test signal to a representation that is similar to the psychophysical representation of audio signals by humans. This is achieved by taking perceptual frequency (Bark) and loudness (Sone) into account. To this end, several stages are included in the algorithm, namely, time alignment, level alignment, time-frequency mapping, frequency warping and compressive loudness scaling [4]. As well, the PESQ algorithm aims to take the severity of effects such as linear filtering and local gain variations into account. This is because these effects, if they are not too severe, may have little perceptual significance. Hence, while minor steady state discrepancies between the original and the test signal are compensated, more sever effects or rapid variations are only partially compensated and will remain to affect the overall perceptual quality. In PESQ, two error parameters are computed in the cognitive model; these are combined to give an objective listening quality score [4].

Wideband PESQ

The wideband extension to PESQ was introduced by ITU-T as P.862.2 standard in 2005 and was amended in 2007. It allows ITU-T Recommendation P.862 to be applied to the evaluation of conditions, such as speech codecs, where the listener uses wideband headphones (In contrast, ITU-T Recommendation P.862 assumes a standard IRS-type narrow-band telephone handset which attenuates strongly below 300 Hz and above 3100 Hz.). The main intention of wideband PESQ is to be used with wideband audio systems (50-7000 Hz), although it can also be applied to narrowband signals [50].

Correlation of PESQ with Reverberation

Very little study has been performed on the correlation of PESQ with reverberation (or lack of reverberation), even though PESQ has been frequently used for evaluation of reverberation. Among the few works which have been carried out on the correlation of PESQ to reverberation, Sharma *et al.* [51] report a very low correlation rate between PESQ prediction and subjective MOS for non-linear distortions such as reverberation. On the other hand, Kokkinakis *et al.* [52] have proposed a modification in the regression model of PESQ score to be adapted to reverberation. In the default scheme, by using three coefficients, the PESQ is calculated as a linear combination of two disturbance indicators as follows

$$\text{PESQ} = a_0 + a_1 D_{ind} + a_2 A_{ind} \tag{2.12}$$

such that

$$a_0 = 4.5, a_1 = -0.1 \text{ and } a_2 = -0.0309$$

where D_{ind} is the average disturbance value and A_{ind} is the average asymmetrical disturbance value. The three parameters are empirically calculated and optimized for speech processed through networks and not for assessing the effects of reverberation (or lack of reverberation) on speech signals [52]. Hence, they propose another combination of the three parameters empirically calculated to better adapt to the task of reverberation calculation. This way, they aim to change the PESQ score calculation to cope with predicting effects of speech coloration, reverberation tail effect, and the overall speech quality in such a manner that is appropriate for reverberation evaluation (for more details and the resulting scheme see [52]).

Nonetheless, this new PESQ scheme has not been standardized or widely accepted and implemented. Due to this fact, and in order to be able to compare the performance of our proposed algorithms with that of similar works, normal PESQ (narrowband and wideband) has been used in this work along with other measures while it has been noted and reminded that PESQ is used for assessing the overall quality of speech signals in a comparative sense.

Perceptual Objective Listening Quality Assessment

Perceptual objective listening quality assessment (POLQA), recommended by ITU-T P.863 standard in 2011, is the successor to PESQ. The main intention of POLQA is for its use with super-wideband systems of today's telecommunication standards [53]. However, researchers are still using the PESQ standard in the very recent works (see for example [24]). In this project, since the signals under test do not exceed the limits of PESQ standard in terms of frequency band, and since the POLQA standard still does not have a guide for implementation, the usage of POLQA has not been followed.

c) Intrusive Channel-based Measures

Direct to Reverberation Ratio

The SRR method introduced earlier in this section was extracted based upon the idea of another measure called direct to reverberation ratio (DRR). The difference between the two measures is that SRR applies to the processed signals while DRR applies to the equalized impulse responses [54].

The DRR is defined as

$$DRR = 10\ \log_{10}\left(\frac{\sum_{n=0}^{n_d} h^2(n)}{\sum_{n=n_d+1}^{\infty} h^2(n)}\right), \qquad [dB] \qquad (2.13)$$

where n_d accounts for the delay of the arrival of the direct component.

2.7. Review of Dereverberation Methods

Dereverberation techniques introduced so far can be classified in different ways. In general, there are only a few recent publications in which a rather broad look into the literature of dereverberation techniques has been given. Dereverberation methods can be split into single-microphone and multi-microphone techniques. Since this work is on single-channel dereverberation, the main focus is on the methods that either have been developed for single-channel dereverberation or have single-channel application addressed in their development specifically. Most of the multi-microphone algorithms cannot be applied to single-channel scenario because they use spatial processing. From another point of view, however, dereverberation methods can be categorized into those primarily focused on coloration and those focused on late reverberation.

Habets [4] classifies dereverberation methods based on whether or not AIR or RIR needs to be estimated. This criterion results in two main categories which he names dereverberation suppression and dereverberation cancellation. Methods in the first category do not estimate the RIR while those in the second category do need to estimate the RIR in order to dereverberate the signal. Habets [4] then splits dereverberation techniques within each category into smaller sub-categories depending on the amount of knowledge about the source or about the acoustic channel that is presumed and used in the method. Fig. 2.4 depicts the two main categories and the sub-categories according to Habets [4]. In the next subsection, the most important and relevant dereverberation techniques classified in the first category are discussed.

2.7.1. Reverberation Suppression

As mentioned before, dereverberation techniques that do not use estimation of the RIR are classified as reverberation suppression techniques. These techniques are in turn classified into sub-categories by considering the amount of knowledge about either the source or the channel, and by the difference in the signal processing techniques that are involved [4].

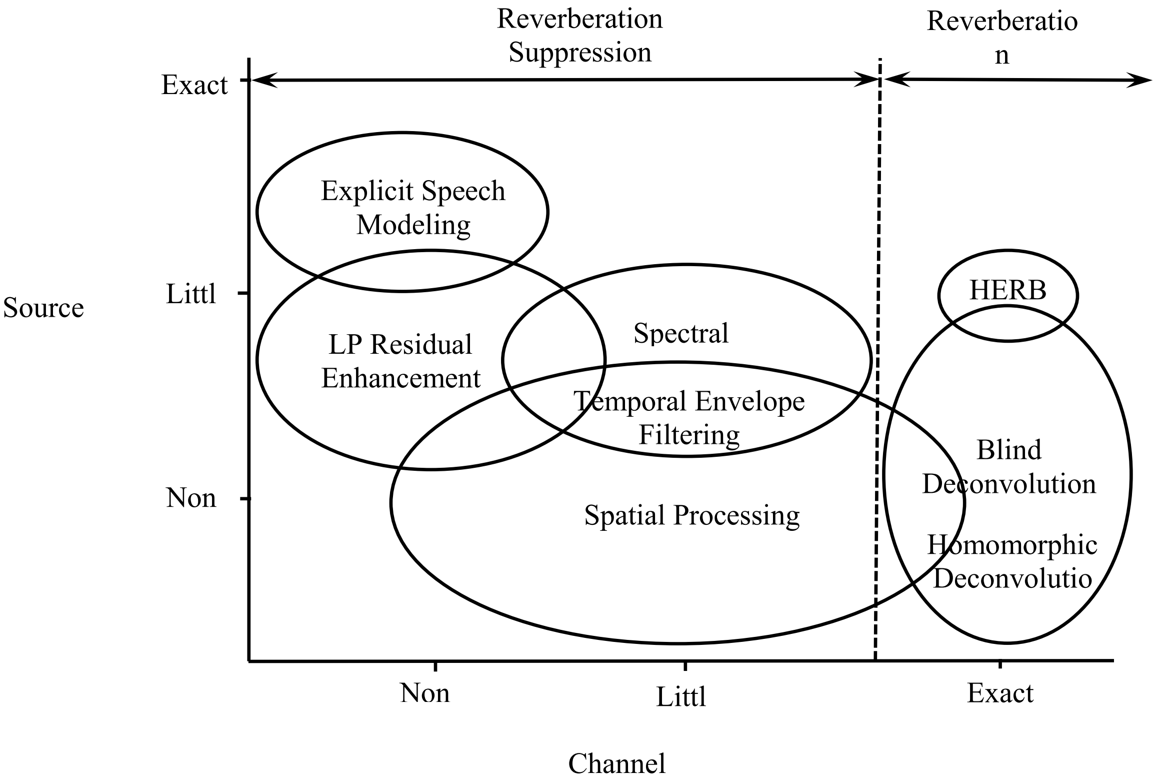


Fig. 2.4. Classification of dereverberation techniques considering the amount of channel and source knowledge used [4].

Explicit Speech Modeling

Some dereverberation methods are based on modeling the speech signal by using the underlying structure of the anechoic speech signal. A dual excitation speech model was proposed by Hardwick in 1992. This model was utilized for speech enhancement purpose in [55]. By adding the effect of pitch variations into the model, it was then complemented to a generalized dual excitation speech model by Yoo [56]. It is remarkable that both of the models mentioned above are based upon the voiced speech segments only.

Brandstein then used the dual excitation model combined with spatial filtering for enhancing reverberant speech in [57]. Later he exploited the generalized dual excitation model in [58].

Attias and Deng utilized probabilistic modeling. In [59], they suggested a unified probabilistic framework for denoising and dereverberation of speech signals. Their proposed framework translates denoising and dereverberation problem to Bayes-optimal signal estimation. The main idea in this method is to pre-train a speech model on a large data set of anechoic speech. This framework is applicable for single- and multi-microphone dereverberation equally well. While their experiments show that optimal Bayesian estimation can outperform standard techniques such as spectral subtraction in terms of noise suppression, unfortunately the dereverberation performance was not evaluated separately. As well, a drawback of this method is that it is strongly dependent on the training of the model [4].

In a more recent work, Nakatani [60] utilized probabilistic features of source signals and room acoustics for single-channel speech dereverberation. The channel was represented by probabilistic density functions (pdf) and the source signals were estimated by maximizing a likelihood function defined based on two types of pdfs. These pdfs were based upon two essential speech signal features, harmonicity and sparseness, while the pdf for the room acoustics is defined based on an inverse filtering operation.

LP-residual Enhancement

Modeling speech as an excitation sequence shaped by a time-varying all-pole filter is a common way to describe the speech signal [46]. The excitation sequence models the unvoiced speech by a random noise sequence and the voiced speech by quasi-periodic pulses. The filter that is used afterwards to shape the speech signal represents the human vocal tract. Figure 2.5 depicts this speech production model. The vocal tract is modelled by an all-pole filter whose coefficients are estimated through linear prediction (LP) analysis of the recorded speech and are called linear prediction coefficients (LPC). In this model, the LP-residual, which represents the excitation sequence, can be obtained by inverse-filtering of the speech signal $s(n)$. The justification of using this inverse-filtering technique is based upon the observation that, in reverberant environment, the LP-residual of voiced speech segments contains the original impulses in addition to several other peaks produced by multi-path reflections. An important assumption made in this technique is that the LPCs are not affected by reverberation. Thus, in general, in this class of techniques, dereverberation is realized by suppressing those peaks in the excitation sequence (LP-residual) which are due to multi-path reflections, and synthesizing the enhanced speech by using the modified LP-residual and the time-varying all-pole filter (the LP-filter) with

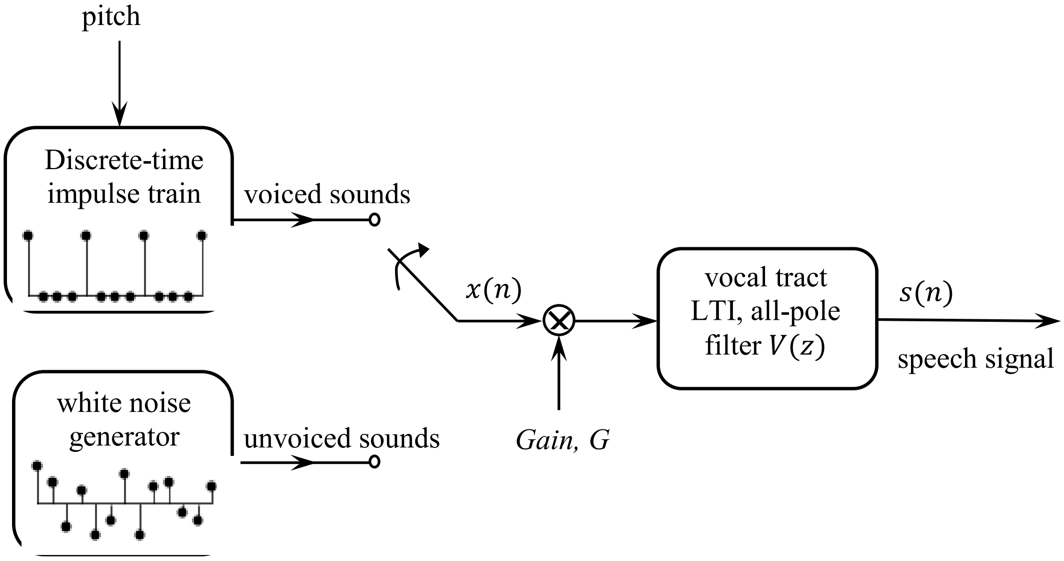


Fig. 2.5. Speech production model [46].

coefficients (LPCs) calculated from the reverberant speech [4].

The general structure of dereverberation by LP-residual enhancement techniques is illustrated in Fig. 2.6. Herein, $x(n)$ represents the samples of the reverberant signal recorded by M microphones at discrete time n. The LPC analysis block stands for the part of the method that estimates the poles of the time-varying all-pole filter shown as $\hat{a}(l)$ (where l represents the frame index) and outputs the error signal, known as the LP-residual signal $\tilde{e}(n)$. Next, based upon some criteria and features depending on the algorithm, the LP residuals are manipulated and the clean LP-residual $\hat{e}(n)$ is estimated. In the next stage, the enhanced speech signal is synthesized by using the estimated poles and the estimated clean LP-residual [4].

Most probably J. B. Allen and F. Haven from Bell Telephone Laboratories Inc. were the first to propose a speech dereverberation algorithm that used the LP-residual enhancement technique in a patent filed in 1974 [61]. This patent addresses both single microphone and multi-microphone scenarios. A detector for separating voiced and unvoiced speech frames, a pitch estimator and a gain estimator are used to synthesize a clean LP residual. Next, they have estimated the vocal tract and used it along with the estimated clean LP-residual to reproduce an estimate of the anechoic

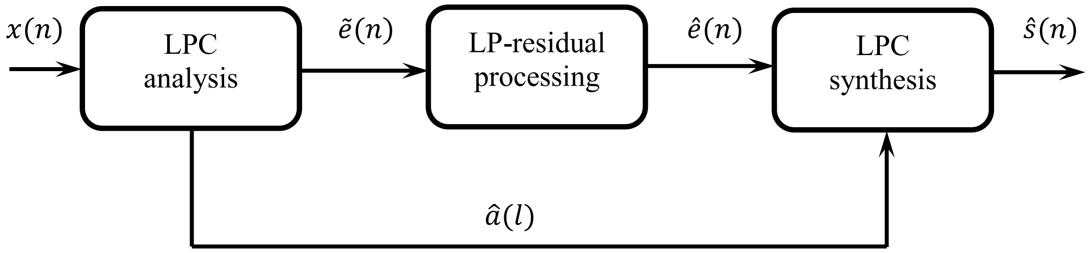


Fig. 2.6. General structure of dereverberation methods that are based on LP-residual enhancement [4].

speech.

In 1999 LP residuals were used by Griebel and Brandstein who proposed a method for multi-channel speech dereverberation by event-based processing of wavelet transform coefficients [62]. The same authors later proposed another multi-channel dereverberation technique in [63] which uses a coarse channel modelling to modify the LP residuals of the channel data.

Yegnanarayana and Murthy developed a single-channel dereverberation technique and comprehensively studied the effects of reverberation on the LP-residual [18, 19]. In their proposed method speech signal is analyzed in short segments (2 ms) to enhance the regions with low SRR. This is based on the observation that in different segments of speech the SRR is different. In their technique, the speech signal is split into three types of regions: low SRR region, high SRR region and regions containing only reverberation components. The LP-residual is modified using a weighting function that assigns different weights to different regions. The time-varying all-pole LP filter then uses the altered LP-residual to form the enhanced speech.

As pointed out earlier, Gillespie *et al.* [21] were the first to perform experiments showing that the kurtosis of the LP residual can be a measure of reverberation. They observed that due to the smearing effect of reverberation on the LP-residual signal, the LP-residual signal becomes less sharp and more Gaussian; hence having lower kurtosis. This technique uses a sub-band adaptive filtering in frequency domain by using a modulated complex lapped transform (MCLT). The subband filters' weights update is performed by maximizing the kurtosis of the LP-residual. As experiments have shown, this method achieves a promising solution to the problem of blind speech dereverberation.

Nonetheless, the calculations of kurtosis and its derivative more or less suffer from instability [64], [65]. To alleviate the instability problem, Tonelli *et al.* [64] proposed a single-microphone dereverberation algorithm based on using a maximum likelihood approach to estimate the inverse-filter. This algorithm was then extended to a multi-microphone dereverberation algorithm in [66].

Yegnanarayana *et al.* [67] exploited the features of the excitation source in speech production model to develop a multi-channel speech enhancement technique. The most important property of the excitation signal is that, in voiced sounds, the strength of excitation is largest around the instant of glottal closure. The strength of excitation was extracted by using the Hilbert envelope of the LP-residual. Then, the Hilbert envelopes of the LP-residual signals from different microphones, after delay compensation, were combined to form a weighting function. The final modified LP-residual was obtained by using this weighting function. By exciting the time-varying all-pole filter with the modified LP-residual the enhanced speech was obtained. Although this method reduces the reverberation effects significantly, it distorts the speech signal to a substantial extent.

Another dereverberation technique based on LP-residual processing was proposed by Gaubitch and Naylor [68]. They enhanced the LP-residual signal from the output of a delay-and-sum beamformer. In contrast to previous algorithms, their method was based on the intention to consider the original structure of the excitation signal. Their method is based on the observation that the LP-residual waveform between adjacent larynx-cycles varies slowly[2]. Therefore, in this method each larynx-cycle is replaced by an average of itself and its' nearest neighbouring cycles. The averaging aims to suppress the additional peaks in the LP-residual introduced by reflections so that the remaining peaks are real peaks produced by the excitation signal. This is based on the observation that in reverberation conditions, in addition to the original excitation impulses, the LP-residual includes several peaks owing to reverberation. In addition, this technique is also based upon the assumption that the calculated LP coefficients of the all-pole filter are unaffected by reverberation. This is while in [69], published one year earlier, they showed that this assumption holds only in a spatially averaged sense and it cannot be guaranteed at a single-point in space for a given room.

In a more recent publication, Gaubitch *et al.* [70] investigated the auto-regressive (AR) (all-pole) modelling of reverberant speech in three different scenarios by using the statistical room acoustic theory. They indicated that, in terms of spatial expectation, the AR parameters calculated from the reverberant speech are approximately equivalent to those of anechoic speech [4]. They showed that this holds for both the single-channel case and the case where the coefficients are jointly computed by a multiple-channel observation. In addition, they showed that the AR coefficients computed at the output of a delay-and-sum beamformer are different from those calculated by using the anechoic speech owing to the spatial correlation between signals from different channels, which depends on the room characteristics and the arrangement of microphones. In general, they indicated that the M-channel joint calculation of the AR coefficients is the preferred option specifically when

[2] The larynx-cycle is the interval of the time from when the glottis opens to when the glottis closes. The length of a larynx-cycle is approximately 20 ms [4].

microphones are closely spaced with a distance of less than 0.3 meter [4]. However, it is notable that all the analyses in these works ([68], [69] and [70]) have been done on a single vowel, i.e., the effects of windowing, self-masking and overlap-masking, have not been taken into the account [1] , [4].

Wu and Wang [22] proposed a two-stage single-channel dereverberation algorithm whose first stage is using the adaptive inverse filtering scheme by kurtosis maximization as proposed by Gillespie *et al.* [21]. In their implementation, however, they utilize the STFT instead of MCLT for transforming to and from the frequency domain. To further improve the dereverberation performance, particularly for long reflections, in the second stage of their proposed algorithm, they have introduced a new and rather complex spectral subtraction scheme to estimate and subtract the reverberation components from the reverberant signal. The resulting two-stage method has been one of the most promising techniques for single-channel speech dereverberation introduced so far and one of the major techniques to compare with in all the works in this area ever since. The same spectral subtraction technique of this algorithm has been used as the second stage of the proposed algorithms of this thesis. Details of this spectral subtraction scheme are explained in Chapter 3, Section 3.2.2.

Nonetheless, this algorithm has two drawbacks. Firstly, it does not obtain good results for rooms with reverberation times of more than 0.5 s. Secondly, background noise conditions have not been considered in their work. These drawbacks have been addressed in the development of our proposed algorithms in Chapter 3.

Later, Kinoshita *et al.* [27] also utilized LP-analysis in their proposed algorithm. This algorithm, in single-channel scenario, consists of pre-whitening and delayed long-term linear prediction on the reverberant speech whose filter coefficients are then used to filter the reverberant speech to obtain an estimation of the reverberation component of the speech. The estimated reverberation component is then subtracted from the reverberant speech in the spectral domain. The output of this analysis is then further enhanced by cepstral mean subtraction, which is not further explained in their work. Although this algorithm might not be considered as one of the major proposed dereverberation algorithms, particularly when it comes to the single-channel case, their scheme of linear prediction has been utilized in the proposed algorithms of this thesis. However, in our work, instead of using the filter coefficients, the LP-residual is utilized to shape an inverse filter based on kurtosis or skewness maximization. Further explanation can be found in Section 3.2 of this thesis.

In a very recent paper, Mosayyebpour *et al.* [26] have proposed another method for inverse-filtering of reverberated speech signal. Their method is also based upon the inverse-filtering scheme proposed by Gillespie *et al.* [21]. However, their algorithm is different in that they utilize the skewness maximization of LP-residual signal rather than kurtosis maximization. They showed that skewness maximization of LP-residual signal, as another measure of non-gaussianity, is superior to kurtosis maximization for the task of dereverberation. Hence, kurtosis as well as skewness will be used in the second phase of the first stage of the proposed algorithms in our work (see Section 3.2.1).

2.8. Summary

This chapter was concerned with providing the theoretical background needed for the study of the dereverberation algorithms proposed in this work. This included the general problem formulation, the concept of AIR and reverberation time and a review of the most relevant reverberation (or dereverberation) measures that have been used in the evaluation of the proposed algorithms in Chapter 4. The last section of the chapter was devoted to a broad classification of dereverberation algorithms and a brief literature review of the most relevant and successful algorithms proposed so far. Explicit speech modelling and LP-residual enhancement as two of the main categories of algorithms classified in reverberation suppression have been reviewed. In particular, as one of the most successful and most relevant category of dereverberation algorithms, LP-residual enhancement based algorithms were reviewed in more detail. It has been shown that although this category of algorithms includes some of the most promising dereverberation methods, there are still some drawbacks which are the focus of the proposed algorithms to be studied in the next chapter.

Chapter 3

Proposed Dereverberation Algorithms

3.1. Introduction

As discussed in detail in Chapters 1 and 2, dereverberation has received a lot of attention in the literature. Most of the focus, however, has been on multi-microphone dereverberation, which is a less challenging problem in general. This is because multi-channel methods allow for both temporal and spatial processing, while single-channel methods are restricted to only temporal processing. The incentive for one-microphone speech enhancement is twofold. First, it is applicable to real world problems such as the processing of telephone speech and audio information retrieval (information extraction from audio signals). Second, one-microphone speech, when moderately reverberated, has the advantage over the multi-microphone case in that it is highly intelligible in monaural listening [22].

Although one-microphone speech dereverberation is more challenging than the multiple-microphones case, a number of algorithms have been proposed in the literature for the former [4], [7], [8], [19], [21]–[23], [26], [27]. Among the single-microphone dereverberation algorithms introduced so far, the one proposed by Wu and Wang [22] is one of the most efficient and most cited algorithms. Although their two-stage algorithm is designed to cancel the short-term and long-term reverberations in the first and second stages, respectively, it is observed that, in the first stage, the inverse filtering based on LP-residuals can be reformed to suppress both the short reflections and long reflections. Also, their method yields satisfactory results only when the reverberation time is short (i.e. less than 0.5 s). Further improvement can be made by using the spectral subtraction in the second stage which in turn suppresses the late reflections in the spectral domain. In a very recent paper [26], the authors have employed skewness maximization of the LP-residuals of the reverberant signal, rather than the kurtosis maximization as was done in [22] as a criterion for adjusting the weights in the inverse filter. However, for speech dereverberation applications, their algorithm is not very effective, especially for long reverberations, as it is based on a single-step LP-residual inverse filtering, which cannot suppress both long and short reverberations at the same time.

Based upon the above observations, in this chapter two new two-stage algorithms are proposed by employing LP-based inverse filtering and spectral subtraction. The first algorithm utilizes the kurtosis maximization for updating the inverse-filter weights, while the second algorithm maximizes the skewness of the LP-residual signal. Except for this difference, and some subsequent minor changes in the parameters, both these algorithms use the same architecture. The algorithms are similar to that by Wu and Wang [22] in that they use normalized higher order moments of LP-residuals for updating the inverse filter weights. However, the proposed algorithms consist of two phases of linear prediction before inverse filtering. In the first phase, the observed signal is whitened by using short-term linear prediction. The second linear prediction phase is a delayed long-term linear prediction as suggested in [27]. These two phases make up the first stage of the proposed algorithms. This is different from the algorithm in [27] in that after applying the delayed long-term linear prediction, the proposed algorithms maximize either the kurtosis or the skewness of the LP-residual for constructing an inverse filter, rather than using the LP-coefficients for estimating late reflections. The second stage of the proposed algorithm is a nonlinear spectral subtraction as proposed by Wu and Wang [22].

3.2. Problem Formulation and Proposed Algorithms

The process of producing a speech sound and the consequent reverberation in a room before the signal is recorded by a microphone is represented by the acoustic system shown in Fig. 3.1. Consistent with the typical speech production modeling, the speech signal is assumed to be produced by a white noise source signal, shown as $u(n)$, shaped by a P-th order FIR filter having a transfer function $A(z)$. The speech signal recorded by the microphone and shown as $x(n)$ is affected by the room impulse response, $b(n)$, which is considered to be invariant in this study. This can be mathematically described as

$$x(n) = \sum_{i=0}^{N} b(i)s(n-i), \tag{3.1}$$

$$= \sum_{l=0}^{T-1} g(l)u(n-l). \tag{3.2}$$

where

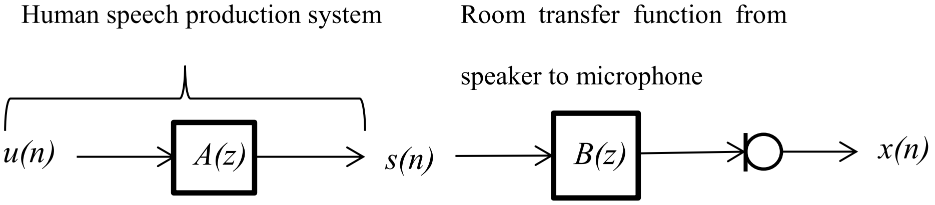

Human speech production system
Room transfer function from speaker to microphone
u(n)
A(z)
s(n)
B(z)
x(n)

Fig. 3.1. Block diagram of the acoustic system.

$$g(l) \triangleq \sum_{k=0}^{P} b(l-k)a(k) \tag{3.3}$$

is the impulse response of the filter obtained by combining the effects of RIR and the human speech production system. Such a filter would produce the recorded speech signal from the white noise sequence $u(n)$. In vector form, this can be formulated as

$$\boldsymbol{x}(n) = \boldsymbol{G}\,\boldsymbol{u}(n) \tag{3.4}$$

where

$$\boldsymbol{x}(n) = [x(n), \quad x(n-1), \dots \;, \quad x(n-N)]^T$$

$$\boldsymbol{g} = [g(0), \quad g(1), \dots \;, g(T-1)]$$

$$\boldsymbol{u}(n) = [u(n),\; u(n-1), \dots \;, u(n-T-N+1)]^T$$

$$\boldsymbol{G} = \begin{bmatrix} g(0) & g(1) & \dots & g(T-1) & 0 & \dots & 0 \\ 0 & g(0) & g(1) & \dots & g(T-1) & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & g(0) & g(1) & \dots & g(T-1) \end{bmatrix}$$

Assuming $\boldsymbol{g}$ and $\boldsymbol{x}(n)$ to be of length T and N respectively, $\boldsymbol{G}$ will be a full row rank matrix of size $(N+1) \times (T+N)$ [27].

The goal of dereverberation in this work is to estimate the clean speech signal, $s(n)$, by observing only the reverberant signal, $x(n)$, without a prior knowledge of $B(z)$.

As mentioned earlier, although the algorithm proposed by Wu and Wang [22] includes spectral subtraction for suppressing the long reflections, it is still not effective enough for suppressing late reverberations and it yields satisfactory results only for RIRs with $RT60$ of less than 0.5 s. This is because in the first stage of their algorithm, the inverse filtering is done on short-LP residuals. The same drawback is found on the inverse-filtering method by Mosayyebpour *et al.* [26]. This inverse filtering method is mostly effective for suppressing colorations (short reverberations) while the main degradation of the quality of the speech signal for both human perception and speech recognition applications is caused by long reverberations. Although the second stage of their algorithm deals with long reflections in the spectral domain, the final results show that further suppression in the time domain is necessary. In other words, the inverse filtering part of their algorithm should be reformed to deal with inverse filtering of both short and long reflections. In order to achieve this goal, in this work, a two-phase linear prediction is introduced before maximizing either the kurtosis or the skewness of the LP-residual signal. The first phase of linear prediction, pre-whitening, accounts for reducing the short-term correlation of a speech signal produced through $A(z)$ and the second phase, delayed long-term linear prediction (DLLP), is to identify the late reverberations.

Although it is out of the scope of this thesis, since there is no constraint regarding the existence of only one observation from the reverberant speech signal, with proper modifications the algorithms should be applicable in the multi-microphone case as well. Clearly, further experiments are needed to prove this claim.

Fig. 3.2. depicts a schematic of the proposed algorithms. The core of the first stage is inverse filtering by maximizing the kurtosis or skewness of LP-residual signal. The signal is passed through two phases of linear prediction before inverse-filtering. In the subsection below, the idea of DLLP and the logic to use it is explained.

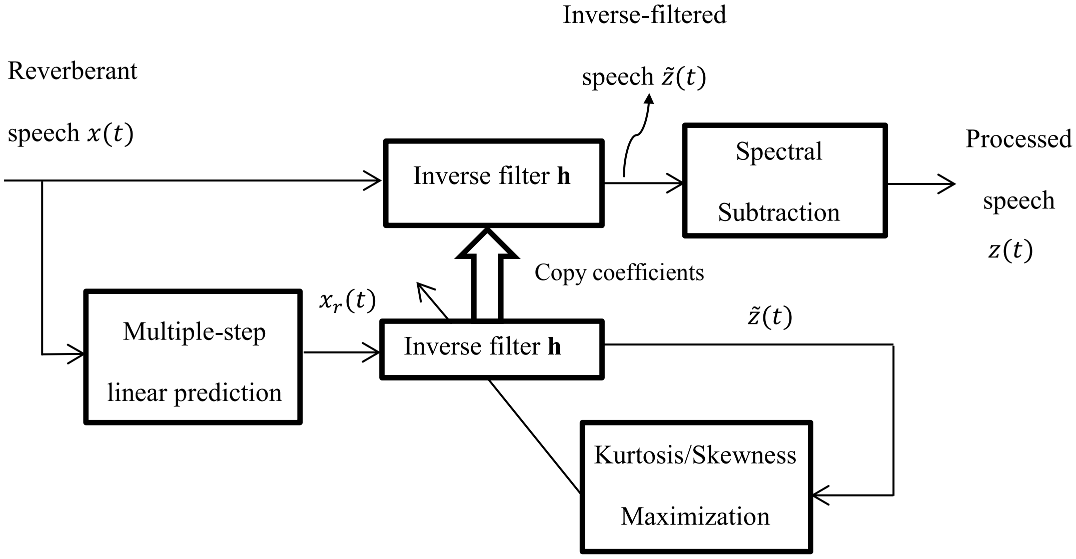


Fig. 3.2. Schematic of the proposed algorithms. Note that multiple-step linear prediction consists of pre-whitening and delayed long-term linear prediction.

Delayed Long-Term Linear Prediction and Pre-Whitening

a) Delayed long-term linear prediction (DLLP)

Delayed long-term linear prediction (DLLP) under the name of *multi-step linear predictor* was used by Gesbert *et al.* [71] for the estimation of a whole impulse response. It was then used by Kinoshita *et al.* [27] for estimating only the late reverberation components to be further used in spectral subtraction. In this work, the same technique is employed to derive LP-residuals rather than LP-filter coefficients. LP-residuals are then used for inverse-filtering by maximization of kurtosis or skewness.

If $x(n)$ is the observed reverberant signal, N is the number of filter coefficients, and D is the step size (the delay) of filtering, the delayed long-term linear prediction is described by

$$x(n) = \sum_{p=0}^{N} w(p)x(n-p-D) + e(n) \tag{3.5}$$

where $w(p)$'s are the filter coefficients and $e(n)$ is the error signal or, alternatively, the LP-residual signal. The conventional linear prediction is a specific case when D is unity. Similar to a normal LP analysis, using the Levinson-Durbin algorithm the mean square energy of the prediction error signal, $e(n)$, is minimized. Using vector notation, when minimizing $e(n)$ one will encounter the following equation, which is the result of Wiener-Hopf equation specialized for delayed linear prediction [27]

$$(E\{\boldsymbol{x}(n-D)\boldsymbol{x}^T(n-D)\})\boldsymbol{w} = E\{\boldsymbol{x}(n-D)x(n)\} \tag{3.6}$$

where

$$\boldsymbol{w} = [w(0), w(1), \dots, w(N-1)]^T.$$

Therefore,

$$\boldsymbol{w} = (E\{\boldsymbol{x}(n-D)\boldsymbol{x}^T(n-D)\})^{-1}E\{\boldsymbol{x}(n-D)x(n)\}. \tag{3.7}$$

It is worth emphasizing that (3.7) is the Wiener-Hopf equation specialized for this case and can be efficiently solved by algorithms such as Levinson-Durbin ([27], [72]), as has been done in the present work.

The first term in (3.7) can be written as

$$E\{\boldsymbol{x}(n-D)\boldsymbol{x}^T(n-D)\} = \boldsymbol{G}E\{\boldsymbol{u}(n-D)\boldsymbol{u}^T(n-D)\}\boldsymbol{G}^T = \sigma_u^2\boldsymbol{G}\boldsymbol{G}^T$$

where $E\{\boldsymbol{u}(n-D)\boldsymbol{u}^T(n-D)\}$, the autocorrelation of white noise, is $\sigma_u^2\boldsymbol{I}$, σ_u^2 being the variance of white noise. As well, the second term in (3.7) can be written as

$$E\{\boldsymbol{x}(n-D)x(n)\} = \boldsymbol{G}E\{\boldsymbol{u}(n-D)\boldsymbol{u}^T(n)\}\boldsymbol{g}^T = \sigma_u^2\boldsymbol{G}\boldsymbol{g}_{late}^T$$

where

$$\boldsymbol{g}_{late} = [g(D),\ \ g(D+1),\ \ \ldots,\ \ g(T-1), 0, \ldots, 0]^T$$

meaning that the first D elements of g are skipped due to the fact that only the rest of them correspond to the part of reverberation that degrades the speech quality [27]. Therefore, we will have

$$\boldsymbol{w} = (\boldsymbol{G}\boldsymbol{G}^T)^{-1}\boldsymbol{G}\boldsymbol{g}_{late} \tag{3.8}$$

By using such a predictor, the estimated power of late reverberations will be

$$E\{(\boldsymbol{x}^T\boldsymbol{w})^2\}$$

$$= \|\boldsymbol{w}^T\boldsymbol{G}\, E\{\boldsymbol{u}(n)\boldsymbol{u}^T(n)\}\boldsymbol{G}^T\boldsymbol{w}\| \tag{3.9}$$

$$= \|\sigma_u^2\boldsymbol{w}^T\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{w}\| \tag{3.10}$$

$$= \|\sigma_u^2\boldsymbol{g}_{late}\boldsymbol{G}^T\ (\boldsymbol{G}\boldsymbol{G}^T)^{-1}\boldsymbol{G}\boldsymbol{g}_{late}\|$$

$$\leq \|\sigma_u^2{\boldsymbol{g}_{late}}^T\|.\ \|\boldsymbol{G}^T(\boldsymbol{G}\boldsymbol{G}^T)^{-1}\boldsymbol{G}\|.\ \|\boldsymbol{g}_{late}\| \tag{3.11}$$

$$= \|\sigma_u\boldsymbol{g}_{late}\|^2. \tag{3.12}$$

Equation (3.10) is obtained by using the fact that $E\{\boldsymbol{u}(n)\boldsymbol{u}^T(n)\} = \sigma_u^2\boldsymbol{I}$, where σ_u^2 represents the variance of $\boldsymbol{u}(n)$. Then, (3.11) is derived by using the Cauchy-Schwartz inequality.

Noting the fact that $\boldsymbol{G}^T\ (\boldsymbol{G}\boldsymbol{G}^T)^{-1}\boldsymbol{G}$ is the norm of a projection matrix and hence, is equal to 1, will result in (3.12) [73]. In addition, (3.12) implies that late reverberations cannot be overestimated [27].

The LP filter order, N, is a large number in the range of several thousands. Therefore, the residual signal each time is computed based on $N + D$ samples [27]. As a result, the LP-residual signal will be able to represent the long-term correlations of the signal. This is in contrast to conventional short-term LP analysis which has been used for short-term dereverberation.

b) Pre-whitening

If the z-domain representation of $g(n)$ and $b(n)$ are $G(z)$ and $B(z)$ respectively, as mentioned before, the long-term delayed LP skips the first D terms of $G(z)$ trying to estimate long reverberations which are harmful to the perceived quality of speech. It should be noted that, as shown in (3.3), $G(z)$ is the product of humans speech production system, $a(z)$, and the room impulse response, $H(z)$. Hence, a bias caused by $A(z)$ exists in estimated late components of $G(z)$. In order to compensate for this bias, pre-whitening by small-order linear prediction is implemented in this work as has been suggested in [27]. However, in this work, the order of pre-whitening was not fixed to 20 taps as suggested in [27], but is adjusted according to the length of the room impulse response. This is due to the fact that the longer the RIR, the longer will be the coloration effect of it on the speech signal. In other words, in this work, pre-whitening compensates for the bias caused by $A(z)$ taking into account its convolution by the room impulse response. Hence, by this modification, the resulting pre-whitening will be more adjusted to the reverberant speech signal under enhancement. Consistent with this theoretical fact, for RIRs with reverberation time equal to 0.9, 0.7 and 0.5, the best pre-whitening short-LP order was empirically chosen equal to 20, 14 and 6 taps, respectively. The final dereverberation result of such an adjustable pre-whitening order scheme proved to be better both by objective and subjective assessments.

Considering the two phases of linear prediction, the term ‘multiple-step linear prediction’ in this work signifies a preprocessing short-order LP followed by the delayed long-order LP.

3.2.1. Inverse Filtering

a) Inverse Filtering by Kurtosis Maximization

As discussed earlier, LP-based inverse filtering has been one of the most powerful dereverberation methods proposed so far. However, LP-based inverse filtering has been mostly used for short-term dereverberation. For suppressing the late reverberations, in some research works, spectral-subtraction-based methods have been used as a second stage after inverse filtering (see for example [22]). The first proposed algorithm in this work consists of two-stages, where the first stage is devoted to inverse filtering of the LP-residual signal by kurtosis maximization and the second stage is assigned to spectral subtraction, see Fig. 3.3. The first stage consists of two phases of linear prediction, namely, pre-whitening and delayed long-term linear prediction (DLLP). The pre-whitening phase is used to suppress the short-term correlation effects; the LP-residual after the DLLP phase represents the long-term

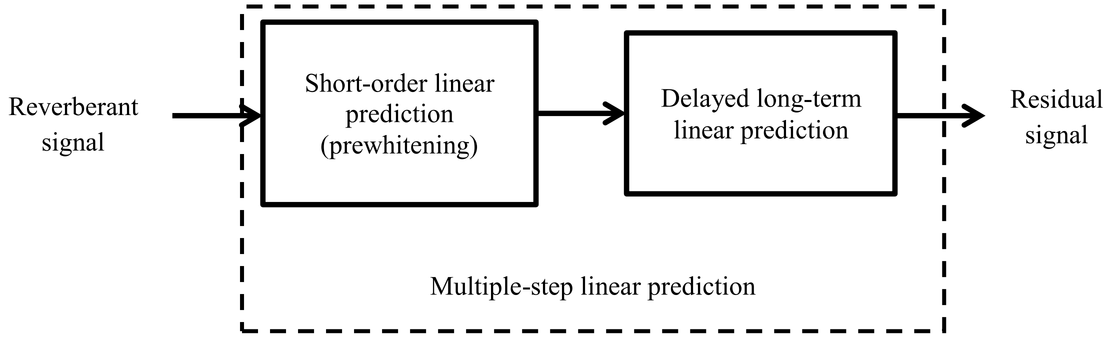

Reverberant signal
Short-order linear prediction (prewhitening)
Delayed long-term linear prediction
Residual signal
Multiple-step linear prediction

Fig. 3.3. Details of Multiple-step linear prediction.

correlations of the reverberant signal. Maximizing the kurtosis of these residuals will be more helpful in suppressing the long reverberations where the actual degrading effect occurs and is the more challenging part of dereverberation.

The LP-based inverse filtering algorithm suggested in [21] estimates the inverse filter of the room impulse response by maximizing the kurtosis of LP-residual signal (i.e. linear prediction error signal). By using the fact that LP-residual of clean signal has a higher kurtosis than that of the reverberant signal, an inverse filter can be estimated using kurtosis maximization of the LP-residual signal. The resulting method is similar to LMS adaptive filtering with the difference that the feedback signal employs kurtosis maximization criterion rather than mean-square error criterion and comparing it to a desired signal. As shown in Fig. 3.2, in this study, the LP-residual is estimated by multiple-step linear prediction of the reverberant speech which includes long-term reverberation effects.

To demonstrate the inverse-filtering we can write

$$\tilde{z}(t) = \boldsymbol{h}\,\widehat{\boldsymbol{x}}_{\boldsymbol{r}}(t) \qquad (3.13)$$

where $\widehat{\boldsymbol{x}}_{\boldsymbol{r}}(t) = [x_r(t-L+1), \ldots, x_r(t-1), x_r(t)]^T$ and $x_r(t)$ is the multiple-step LP-residual of the reverberant speech, $\boldsymbol{h}$ is the inverse filter and $\tilde{z}(t)$ is the inverse-filtered signal. In the feedback path, the kurtosis of $\tilde{z}(t)$, is maximized and the inverse filter is modified accordingly. The kurtosis of the residual signal is given by

$$J_1(t) = \frac{E\{\tilde{z}^4(t)\}}{E^2\{\tilde{z}^2(t)\}} - 3$$

As proved in [21], by taking the gradient of the kurtosis with respect to the inverse filter we obtain

$$\frac{\partial J_1(t)}{\partial \boldsymbol{h}(t)} = \frac{4(E\{\tilde{z}^2(t)\}E\{\tilde{z}^3(t)\widehat{\boldsymbol{x}}_{\boldsymbol{r}}(t)\} - E\{\tilde{z}^4(t)\}E\{\tilde{z}(t)\widehat{\boldsymbol{x}}_{\boldsymbol{r}}(t)\})}{E^3\{\tilde{z}^2(t)\}}$$

Similar to [74] the gradient could be approximated by

$$\frac{\partial J_1(t)}{\partial \boldsymbol{h}(t)} \approx \left(\frac{4\big((E\{\tilde{z}^2(t)\}\tilde{z}^2(t) - E\{\tilde{z}^4(t)\})\tilde{z}(t)\big)}{E^3\{\tilde{z}^2(t)\}}\right)\widehat{\boldsymbol{x}}_{\boldsymbol{r}}(t) = f_1(t)\widehat{\boldsymbol{x}}_{\boldsymbol{r}}(t)$$

where $f(t)$ is referred to as the feedback function controlling the coefficient updates of the inverse filter. In order to do the inverse filtering adaptively, $E\{\tilde{z}^2(t)\}$ and $E\{\tilde{z}^4(t)\}$ are calculated recursively by

$$E\{\tilde{z}^2(t)\} = \beta\, E\{\tilde{z}^2(t-1)\} + (1-\beta)\tilde{z}^2(t), \quad and$$

$$E\{\tilde{z}^4(t)\} = \beta\, E\{\tilde{z}^4(t-1)\} + (1-\beta)\tilde{z}^4(t).$$

where the parameter β controls the smoothness of the moment estimates. Consequently, the adaptive inverse filter that maximizes the kurtosis of the input signal can be described by the following weight update equation which represents a time-domain adaptive filter implementation of the method [21].

$$\boldsymbol{h}(t+1) = \boldsymbol{h}(t) + \mu\, f(t)\widehat{\boldsymbol{x}}_{\boldsymbol{r}}(t) \tag{3.14}$$

where

$$f_1(t) = \frac{4\big((E\{\tilde{z}^2(t)\}\tilde{z}^2(t) - E\{\tilde{z}^4(t)\})\tilde{z}(t)\big)}{E^3\{\tilde{z}^2(t)\}} \tag{3.15}$$

and μ adjusts the learning rate for the weight update of the inverse filter. However, according to Haykin [75] and as reflected also in [21] and [22], the time domain implementation of such an adaptive filter is not recommended because of the large variations in the eigenvectors of the autocorrelation matrices of the input signal which can result in very slow or no convergence. As a result, a block-frequency domain implementation is adopted in this work consistent with [21] and [22]. Herein, a frame-by-frame processing of the signal is performed in the frequency domain by using the STFT and its inverse for transforming to and from the frequency domain. This is in contrast to the original implementation of the technique in [21], which utilizes the modulated complex lapped transform (MCLT) and its inverse for this task. The block length for FFT is chosen to be the same as the filter length. In the frequency domain, the inverse filtering equations will be

$$\acute{\boldsymbol{H}}(n+1) = \boldsymbol{H}(n) + \frac{\mu}{M}\sum_{m=1}^{M} \boldsymbol{F}(m)\boldsymbol{X}_r^*(m) \tag{3.16}$$

$$\boldsymbol{H}(n+1) = \frac{\acute{\boldsymbol{H}}(n+1)}{\left|\acute{\boldsymbol{H}}(n+1)\right|} \tag{3.17}$$

where $\boldsymbol{F}(m)$ and $\boldsymbol{X}_r(m)$ are the FFT of $f(t)$ and $\hat{\boldsymbol{x}}_r(t)$ for the *m*-th block respectively, the superscript $*$ denotes complex conjugate, $\boldsymbol{H}(n)$ is the FFT of $\boldsymbol{h}$ at *n*th iteration, and M is the total number of blocks (i.e. frames here because each frame is transferred to one block in frequency domain). The second equation above, (3.17), is to normalize the inverse-filter weights so as to prevent the blowing up of the speech volume in the output. The inverse-filtered speech is obtained by convolving the reverberant speech with the adjusted inverse filter in the time domain.

Henceforth, this inverse-filtering method, along with the spectral subtraction as the second stage, is referred to as Algorithm 1.

Next, inverse filtering by skewness maximization is described.

b) *Inverse Filtering by Skewness Maximization*

As implied earlier, Mosayyebpour *et al.* [26] observed that maximizing the skewness of sufficiently long LP-residuals can be a more efficient method for dereverberation with some advantages both in effectiveness and robustness. In this work, as a second technique, the skewness of the LP-residuals is maximized to update the weights of the inverse filter.

The skewness is defined as

$$J_2(t) = \frac{E\{\tilde{z}^3(t)\}}{E^{\frac{3}{2}}\{\tilde{z}^2(t)\}}$$

Hence, by taking the gradient of skewness with respect to the inverse filter we will have

$$\frac{\partial J_2(t)}{\partial \boldsymbol{h}(t)} = \frac{3(E\{\tilde{z}^2(t)\}E\{\tilde{z}^2(t)\hat{\boldsymbol{x}}_{\boldsymbol{r}}(t)\} - E\{\tilde{z}^3(t)\}E\{\tilde{z}(t)\hat{\boldsymbol{x}}_{\boldsymbol{r}}(t)\})}{E^{\frac{5}{2}}\{\tilde{z}^2(t)\}}$$

which with the same approximation as for the kurtosis case, we obtain

$$\frac{\partial J_2(t)}{\partial \boldsymbol{h}(t)} \approx \left(\frac{3\big((E\{\tilde{z}^2(t)\}\tilde{z}^2(t) - E\{\tilde{z}^3(t)\})\tilde{z}(t)\big)}{E^{\frac{5}{2}}\{\tilde{z}^2(t)\}}\right)\hat{\boldsymbol{x}}_{\boldsymbol{r}}(t) = f_2(t)\hat{\boldsymbol{x}}_{\boldsymbol{r}}(t) \qquad (3.18)$$

where with the same weight update equation, (3.14), we will have

$$f_2(t) = \frac{3\big((E\{\tilde{z}^2(t)\}\tilde{z}^2(t) - E\{\tilde{z}^3(t)\})\tilde{z}(t)\big)}{E^{\frac{5}{2}}\{\tilde{z}^2(t)\}} \qquad (3.19)$$

Here again the inverse-filtering and skewness maximization are performed in the frequency domain; therefore, (3.16) and (3.17) hold. The only difference is that in skewness maximization the length of the inverse filter and the parameter D (delay of the DLLP) are different.

Unlike kurtosis maximization, skewness maximization is sensitive to the inverse-filter length. In other words, for longer reverberations, longer inverse-filter length should be adopted. By investigating this effect, Mosayyebpour *et al.* [26] have found the optimum inverse-filter length for different RIR lengths for satisfactory performance with the lowest computation. The same general rule is applied in this work. This means for longer RIR a longer inverse-filter length is chosen. However, based on our experiments, the optimal number of taps in this work range from 1024 taps for RIR with $RT60 = 0.5\ s$ and $RT60 = 0.7\ s$ to 2048 taps for $RT60 = 0.9\ s$. One source of discrepancy of the inverse-filter lengths in our work with those of Mosayyebpour *et al.* [26], could be due to the differences in the implementation of simulated RIRs.

Hereafter, this inverse-filtering method, along with spectral subtraction as the second stage, is referred to as Algorithm 2.

It may be mentioned that the delayed long-term LP increases the execution time of the algorithms due to the delay D and due to the fact that calculations of the long-term correlations are performed on large frames of length N. In addition, prewhitening, as another phase of linear prediction, is expected to add to the execution time of the algorithms as compared to the method of Wu and Wang [22] and that of Mosayyebpour *et al.* [26].

In the next section, the second stage of the algorithm, spectral subtraction, is described

3.2.2. Spectral Subtraction

As the second part of the algorithms, a nonlinear spectral subtraction stage, similar to that in [22], is implemented. This is to further suppress the long reverberations in the observed signal.

As mentioned before, an impulse response, like the one shown in Fig. 3.4, consists of two parts: early and late impulses. The late impulses, which represent the effects of late reverberations in a room impulse response, have damaging effect on the quality of inverse-filtered speech. Thus, it is helpful to spectrally estimate the late reflections and subtract them from the reverberant speech in the spectral domain. It is notable that although in these algorithms the inverse-filtered speech has been derived using the long-term linear prediction, which in turn alleviates the problem of late reflections more than in conventional linear prediction, spectral subtraction can still enhance the

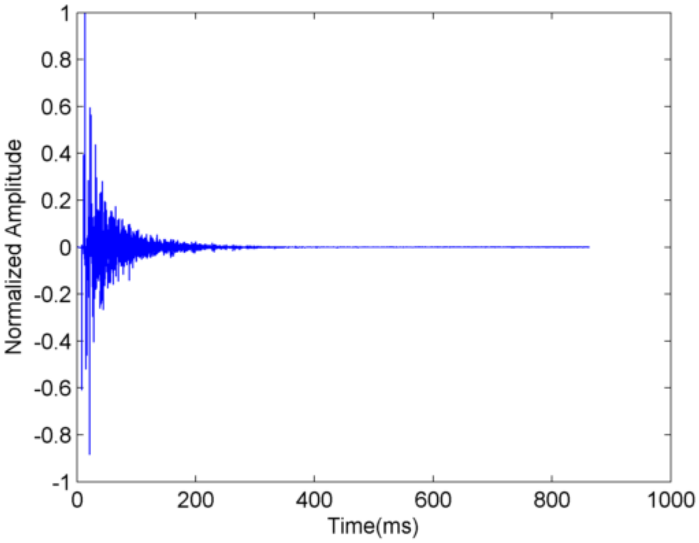


Fig. 3.4. RIR with RT60=0.5 s simulated by image method.

quality of speech signal further, since it does the dereverberation in spectral domain rather than in time domain [22].

In order to suppress late reverberations a number of methods have been introduced. Amongst these, several algorithms have tried to spectrally subtract the estimated spectrum of late reflections from that of the reverberant signal. However, in general, the differences in proposed algorithms have been twofold:

1. The way the spectra of long reverberations are estimated.
2. The way the spectral subtraction is performed on the spectra including linear or nonlinear subtraction, thresholds and constraints.

As an example, Kinoshita *et al.* [27] developed a dereverberation method based on spectrally subtracting the late reverberations from the reverberant signal. They used normal spectral subtraction, but employed a different technique to estimate the long reverberations. By using multiple-step linear prediction, including pre-whitening and delayed long-term linear prediction, they obtained a set of appropriate filter coefficients to be applied to the reverberant signal and estimated the late reverberations. Afterwards, they employed a simple spectral subtraction to subtract the long reverberation components from the reverberant signal.

Wu and Wang [22], on the other hand, use a spectral subtraction that estimates the late-impulse components by a Rayleigh function and subtracts them in the spectral domain considering a specific time lag and also different thresholds. This is a rather complex, yet promising spectral subtraction method. The method is used in spectral subtraction phase of our algorithm.

The method is based on the fact that effects of late impulse components result in the smoothing of the signal spectrum in time. Therefore, it is assumed that the power spectra of late impulses can be modeled as smoothed form of the power spectrum of the inverse-filtered speech which is shifted by a specific time lag [22]. This can be shown as

$$|S_l(k;i)|^2 = \gamma\omega(i-\rho) * |S_z(k;i)|^2 \tag{3.20}$$

where $|S_l(k;i)|^2$ and $|S_z(k;i)|$, respectively, represent the short term power-spectra of the late impulse components and that of the inverse-filtered speech, index k stands for the frequency bin and index i refers to the time frame. The right side is convolution of the smoothing function, $\omega(i)$, with the spectrum of the inverse-filtered speech at the time frame i. The spectrum is magnitude squared of the STFT of the signals. Hamming windows of length 16 ms with 8-ms overlap are used for STFT. The shift of ρ in the smoothing function indicates the delay of the late impulse components. Disregarding the reverberation characteristics, and considering only the speech characteristics in general, the border of distinction between early and late reverberations in speech is commonly set at 50 ms. This time interval translates to 7 frames for the windowing in our work. Consequently, $\rho = 7$ is used here. In addition, γ is a scaling factor controlling the relative strength of late impulses and empirically is set to 0.32. A detailed analysis of the effect of changing the scaling factor, γ, and its relation to reverberation time is given in [22], although it has been concluded that its detailed values do not matter.

Due to the shape of the impulse response, an asymmetrical smoothing function, namely, Rayleigh distribution, is chosen for estimating late impulses as follows

$$\begin{cases} \omega(i) = \frac{i+a}{a^2} \exp\left(\frac{-(i+a)^2}{2a^2}\right), & if\ i > -a \\ \omega(i) = 0, & otherwise. \end{cases} \tag{3.21}$$

The smoothing function is illustrated in Fig. 3.5. The overall spread of the function is controlled by parameter a, which is supposed to be smaller than ρ and in this implementation is set to 4 empirically.

Owing to the long-term uncorrelation of speech signal, early and long reflections can be assumed to be almost uncorrelated. Hence, the power spectrum of the early-impulse components can be estimated by subtracting the power spectrum of the late-impulse components from that of the inverse-filtered speech [22]. Moreover, this power spectrum can be used as an approximation of the power spectrum of the original clean speech.

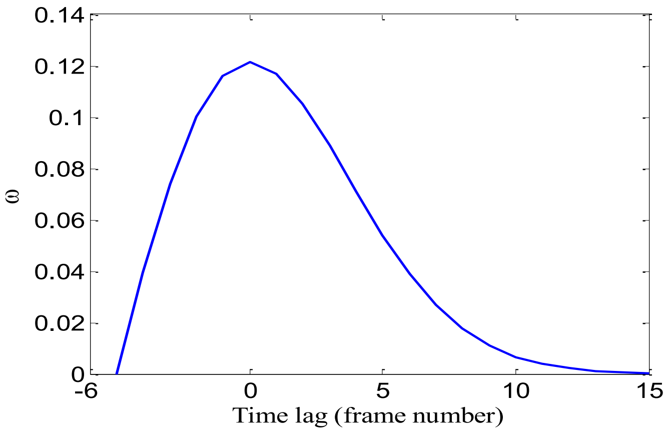


Fig. 3.5. The smoothing function corresponding to equation (3.21) for a = 5 [22].

In this algorithm, spectral subtraction is performed according to the following equation

$$\left|S_p(k;i)\right|^2 = |S_z(k;i)|^2 \max\left[\frac{|S_z(k;i)|^2 - \gamma\omega(i-\rho) * |S_z(k;i)|^2}{|S_z(k;i)|^2}, \varepsilon\right] \tag{3.22}$$

where $\varepsilon = 0.001$ is the threshold of the attenuation of the late components corresponding to a maximum of 30 dB, and $\left|S_p(k;i)\right|^2$and $|S_z(k;i)|^2$ represent the short-term spectra of processed speech and inverse-filtered speech, respectively.

Another important part of the employed spectral subtraction method, is the detection of silent gaps in the speech signal and further suppression of reverberations in such frames. Therefore, the inverse-filtered signal is first normalized so that the maximum energy of frames is unity. Then, if a frame's energy level is lower than a predefined threshold, ϑ_1, the frame is considered to be a candidate for a silent frame. Next, for such frames a second condition is checked. If the proportion of the energy value of the inverse-filtered speech to the energy value of processed speech, $E_z(i)/E_p(i)$, is greater than a second threshold, ϑ_2, the frame is identified as a silent frame for which all the frequency bins are attenuated by 30 dB. The silent frame detection rules are as follows

$$frame\ i\ is\ silent\ if \begin{cases} 1.\ E_z(i) < \vartheta_1, & \vartheta_1 = 0.0125 \\ 2. \dfrac{E_z(i)}{E_p(i)} > \vartheta_2, & \vartheta_2 = 5 \end{cases}$$

In our implementation of spectral subtraction, except for $a = 4$, which is in accordance with the MATLAB source code of the algorithm, the value of all other parameters, including $\rho, \gamma, \vartheta_1, \vartheta_2$ are identical to the original values suggested in

[22][3].

3.3. Summary

In this chapter two new algorithms have been proposed for single-channel speech dereverberation. The proposed algorithms consist of two stages. The first stage of the algorithms has two phases, pre-whitening followed by delayed long-term linear prediction. In Algorithm 1, the kurtosis of the resulting LP-residual signal was maximized to form the inverse filter; whereas in Algorithm 2, the skewness of the signal was employed rather than the kurtosis. The second stage of the algorithms is a nonlinear spectral subtraction, as proposed by Wu and Wang [22]. Based upon the theoretical analysis given, the resulting algorithms should be capable of removing both the short and long reflections more effectively for RIRs with short and long reverberation times. Also, it is expected that, due to the prewhitening and the spectral subtraction utilized in the proposed algorithms, they would be relatively more robust to the background noise.

In the next chapter, details of the experiments conducted to assess the performance of the proposed algorithms are given. Also, the results are compared to those of Wu and Wang [22] and Mosayyebpour *et al.* [26], which are among the most relevant and most promising single-channel dereverberation algorithms at present.

[3] The implementation of spectral subtraction has been according to the available MATLAB code associated with [22] available at http://www.cse.ohio-state.edu/~dwang/pnl/shareware/wu-taslp06/.

Chapter 4

Performance of Proposed Algorithms

4.1. Introduction

This chapter is concerned with the performance evaluation of the two dereverberation algorithms proposed in Chapter 3 and comparison of their performance with that of two of the most successful existing single-channel dereverberation algorithms. First, the experimental setup and the parameters used in the implementation of the proposed and the existing algorithms are described. Then, the results of the algorithms obtained using qualitative and quantitative measures are discussed. The measures chosen are based on the type of the algorithms and are consistent with similar works in the literature.

4.2. Experimental Setup and Simulation Parameters

In this study, the sampling frequency for both the speech signals and the room impulse responses is chosen to be $f_s = 16\,kHz$. Delayed long-term LP filter uses a filter of length $N = 6000$ taps. The delay factor for the first proposed algorithm (Algorithm 1) is $D = 480$ samples. However, for the second proposed algorithm (Algorithm 2), it was observed that a reduced delay of $D = 80$ is more effective. In contrast to the typically-fixed short-order LP in previous works, the short-order LP in the pre-whitening phase of this work has a varying filter order of 20, 14, and 6 taps for RIRs with reverberation times of 0.9, 0.7, and 0.5 s, respectively. The reason for choosing such a variable pre-whitening order has been explained in part b of Section 3.2.1. Simulations are performed on the TIMIT speech database including 32 speakers from 8 different dialects of English language[4]. The length of each utterance is about two seconds. For the inverse filtering part, we choose $\mu = 3 \times 10^{-9}$, $\beta = 0.995$ and the number of iterations to be 1000. These are identical to the parameters used in the implementation of Wu and Wang [22] and Mosayyebpour *et al.* [26]. The room impulse responses are generated based on the image method introduced by Allen and Barkley [76]. In our study, the MATLAB implementation of this method by Lehmann is used[5]. An example of the RIR of the simulated rooms with reverberation time of $RT60 = 0.5\ s$ was depicted in Fig. 3.4. The simulated room is of dimensions (6×4×3) meters, with the microphone positioned at (4, 1, 2) and the source positioned at (2, 3, 1.5). The reflection coefficients of the walls are [0.95, 0.95, 0.85, 0.85, 0.80, 0.80]. Two more rooms with $RT60 = 0.7\ s$ and $RT60 = 0.9\ s$ are simulated. By using the Schroeder method[6], the $RT60$ values are validated after simulation and some necessary minor modifications are performed.

The proposed algorithms are compared with those of Wu and Wang [22] and Mosayyebpour *et al.* [26], which are the two important existing algorithms for single-channel dereverberation and inverse-filtering of speech, respectively. As the code[7] of the algorithm of [22] is available, this algorithm is implemented in a way identical to

[4] The TIMIT database is a licensed database available for purchase at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.

[5] http://www.mathworks.com/matlabcentral/fileexchange/20962-image-source-method-for-room-impulse-response-simulation-room-acoustics.

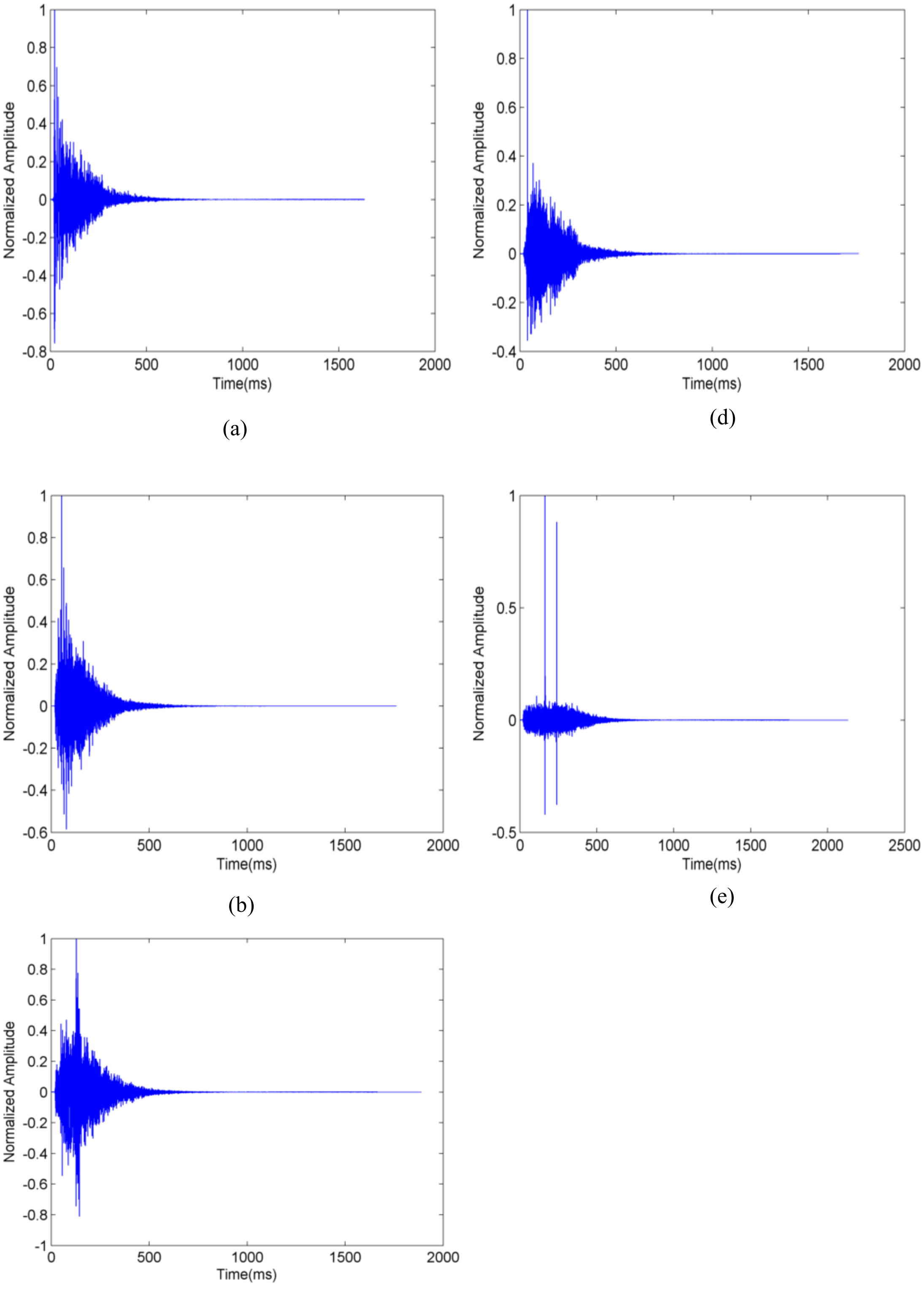
[6] http://www.mathworks.com/matlabcentral/fileexchange/35740-blind-reverberation-time-estimation/content/utilities/RT_schroeder.m.

[7] The source code is available at http://www.cse.ohio-state.edu/~dwang/pnl/shareware/wu-taslp06/

the source code available. The algorithm of [26] is implemented according to the information, the parameters and all the considerations regarding the implementation issues as reported in [26]. These considerations include, for example, the inverse filter length and the data size.

4.3. Equalized Impulse Responses and Energy Decay Curves

Fig. 4.1 shows the original RIR with $RT60 = 0.9\ s$ along with its equalized versions by the proposed Algorithms 1, and 2, and those of Wu and Wang [22] and Mosayyebpour *et al.* [26]. The equalized RIRs are the results of the convolution of the impulse response of the derived inverse filter with the original RIR. This is to evaluate the performance of the inverse-filtering stage of the algorithms. As can be seen from the figure, associated with a long RIR of $RT60 = 0.9\ s$, the inverse filtering of the proposed Algorithms 1 and 2 demonstrate a superior capability in suppressing the late impulse components as compared to the algorithm of Wu and Wang [22]. It should be pointed out that the late impulse components are more deleterious to the quality of speech signal both for perception and for automatic speech recognition systems. On the other hand, at first glance, the equalized RIR by the method of Mosayyebpour *et al.* [26] seems to be more succesful in suppressing the short and long impulse components. However; it has two drawbacks. First, the equalized RIR by the method of Mosayyebpour *et al.* [26] has two rather distant peak impulses. This, as will be examined later by using a reverberation time estimation method, increases the reverberation time of the RIR. Second, this equalized RIR does not preserve the overall shape of the original RIR. This results in a speech signal that does not sound natural.



(a)

(d)

(b)

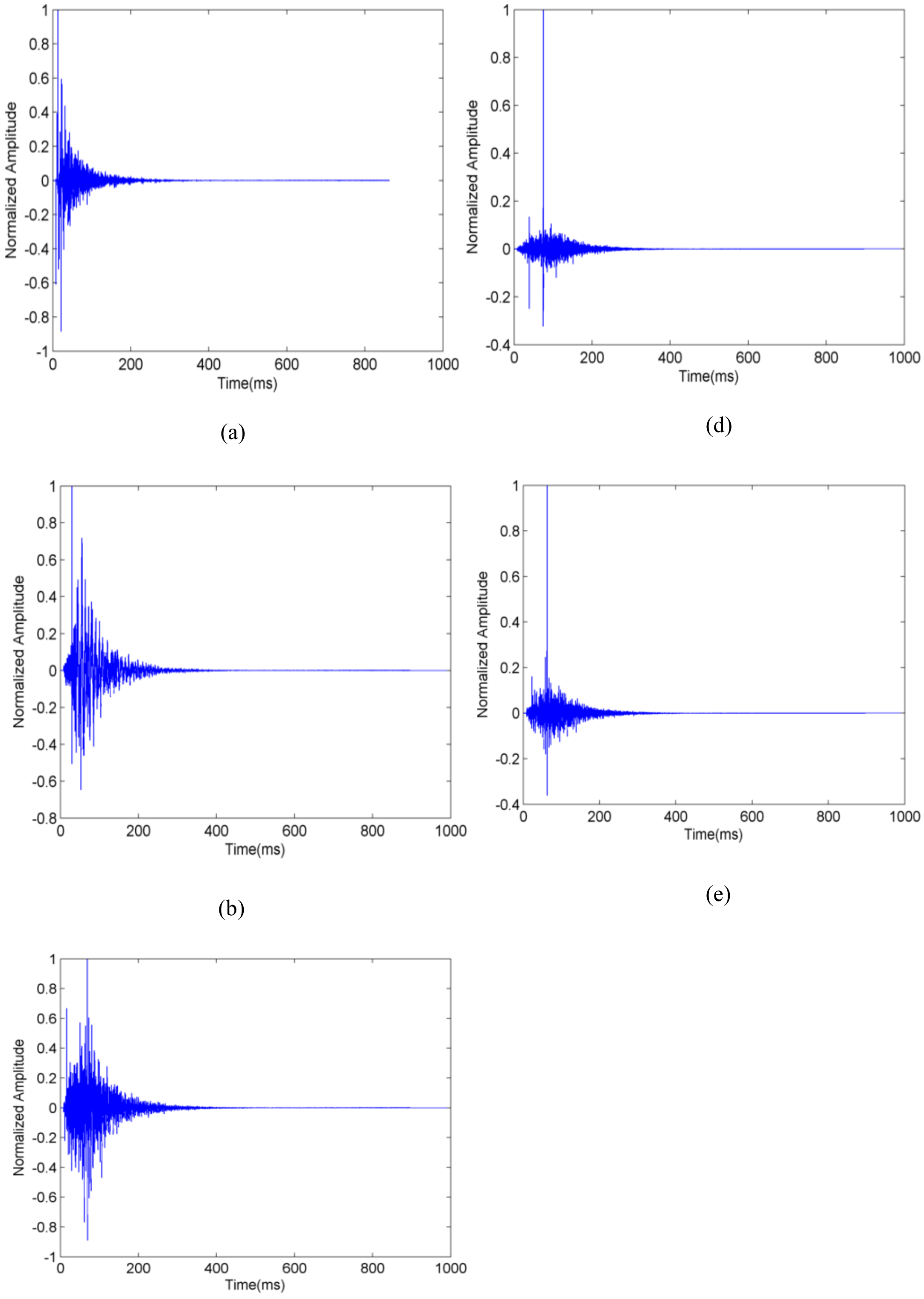
(e)

(c)

Fig. 4.1. (a). Room Impulse response with RT60=0.9 s, (b) the same RIR equalized by Algorithm 1, (c) the same RIR equalized by Algorithm 2, (d) the same RIR equalized by the algorithm of Wu and Wang [22], and (e) the same RIR equalized by the algorithm of Mosayyebpour *et al.* [26].

Fig. 4.2 depicts the original RIR with $RT60 = 0.5\ s$ along with its equalized versions by Algorithm 1, Algorithm 2, the method of Wu and Wang [22] and that of Mosayyebpour *et al.* [26]. In this figure, the difference between the algorithms is more clear. Here, the methods of Wu andWang [22] and Mosayyebpour *et al.* [26] suppress almost all the impulses except for one impulse related to the direct path. In contrast, in Algorithm 1, although equalization has removed some mid to late impulses, the overal pattern of the RIR is not changed. This is helpful for maintaining the overal perceived sound quality and naturalness of the speech. As will be examined shortly, as compared to the the existing algorithms under experimentation here, the equqlized RIRs by the proposed algorithms have less or equal reverberation times while preserving the overal pattern of the RIR.

In order to compute the reverberation times of the equalized RIRs, the Schroeder method is used in our work. This mehod, whose reference of the MATLAB code was given before in Section 4.2, uses the energy curve of the RIRs in order to calculate the reverberation time. Fig. 4.3 illustrates the energy curve of the original RIR with $RT60 = 0.9\ s$ along with its equalized versions by different algorithms. A close look at Fig. 4.3, and considering the fact that the x axis is not of the exact same length in time in different graphs, indicates that all the equalized RIRs experience more energy decay than the original RIR. The shape of the energy curves follows and confirms the shape of the impulse responses. For instance, in Fig 4.3 (e), the energy curve includes two drastic drops which correspond to the two peak impulses in the impulse response shown in Fig 4.1 (e). As well, it can be detected that the equalized RIRs by the proposed algorithms experience a little bit of more decay at the end as compared to the other two algorithms. Also, Fig. 4.4 depicts the energy decay curves of the original RIR with $RT60 = 0.5\,s$ along with its equalized versions by different algorithms. The comments made for Fig. 4.3 hold true for Fig. 4.4 also. By applying the Schroeder method to these energy decay curves, the reverberation time values of the equalized RIRs are calculated. For the purpose of simplicity and clarity of the figure, the details related to the calculation of $RT60$ valuses by using the Schroeder method are not shown in the figure. The difference in the reverberation times between the equalized RIRs is only clear when looking at Table 4.1, which includes the estimated $RT60$ values for the same impulse responses. Comparing the estimated $RT60$ values of the table confirms the superior capability of Algorithms 1 and 2 to that of Wu and Wang [22] and Mosayyebpour *et al.* [26] in equalizing the RIR. For both RIR with $RT60 = 0.5\,s$ and RIR with $RT60 = 0.9\,s$ the two proposed algorithms result in equalized RIRs with $RT60$ values less than that of the original RIR and those of the two benchmark algorithms. This, in turn, means the inverse filters of our algorithms are more successful in cancelling the reverberations in the speech signal.



(a)

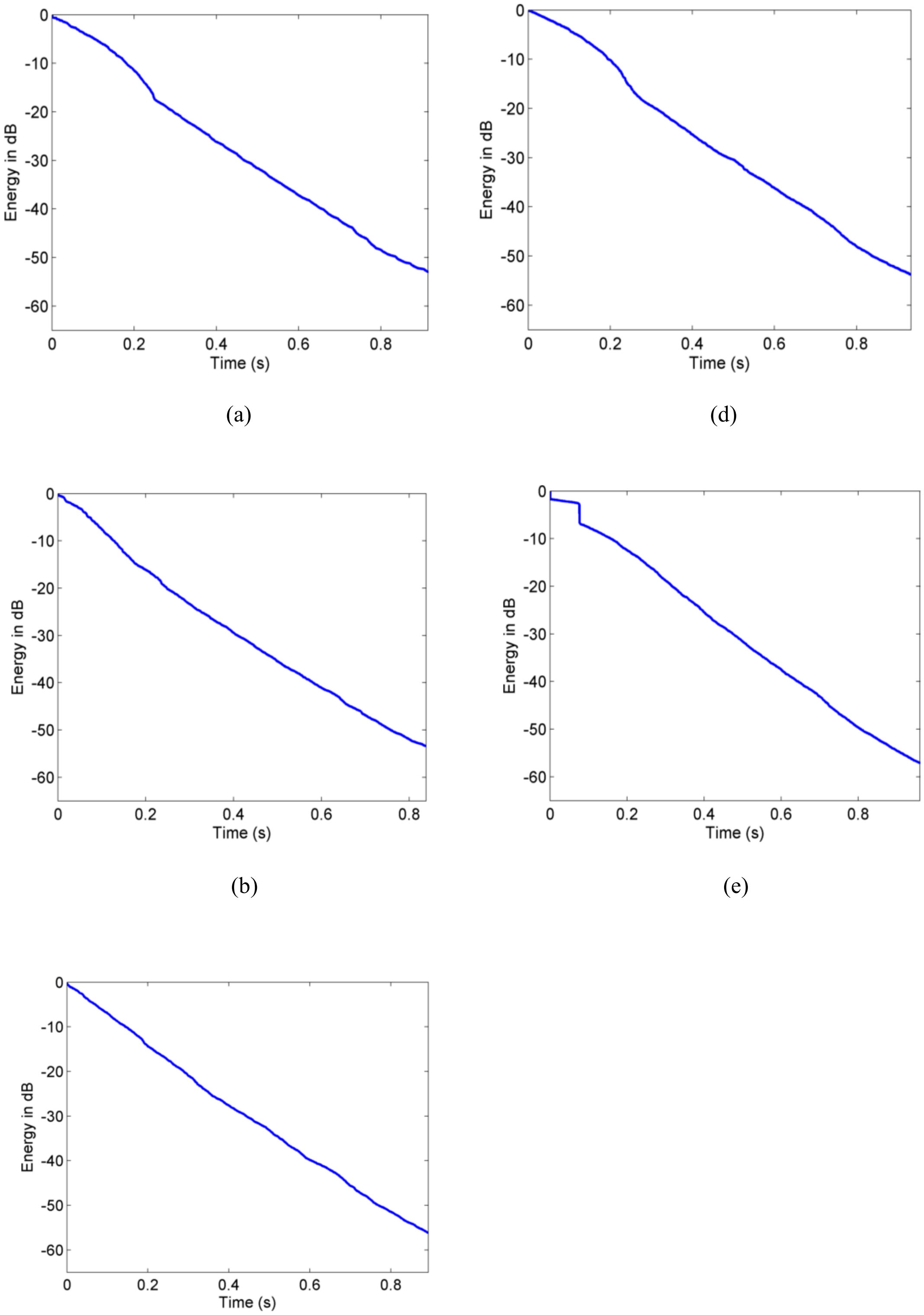
(d)

(b)

(e)

(c)

Fig. 4.2. (a). Room Impulse response with RT60=0.5 s, (b) the same RIR equalized by Algorithm 1, (c) the same RIR equalized by Algorithm 2, (d) the same RIR equalized by the algorithm of Wu and Wang [22], and (e) the same RIR equalized by the algorithm of Mosayyebpour *et al.* [26].



(a)

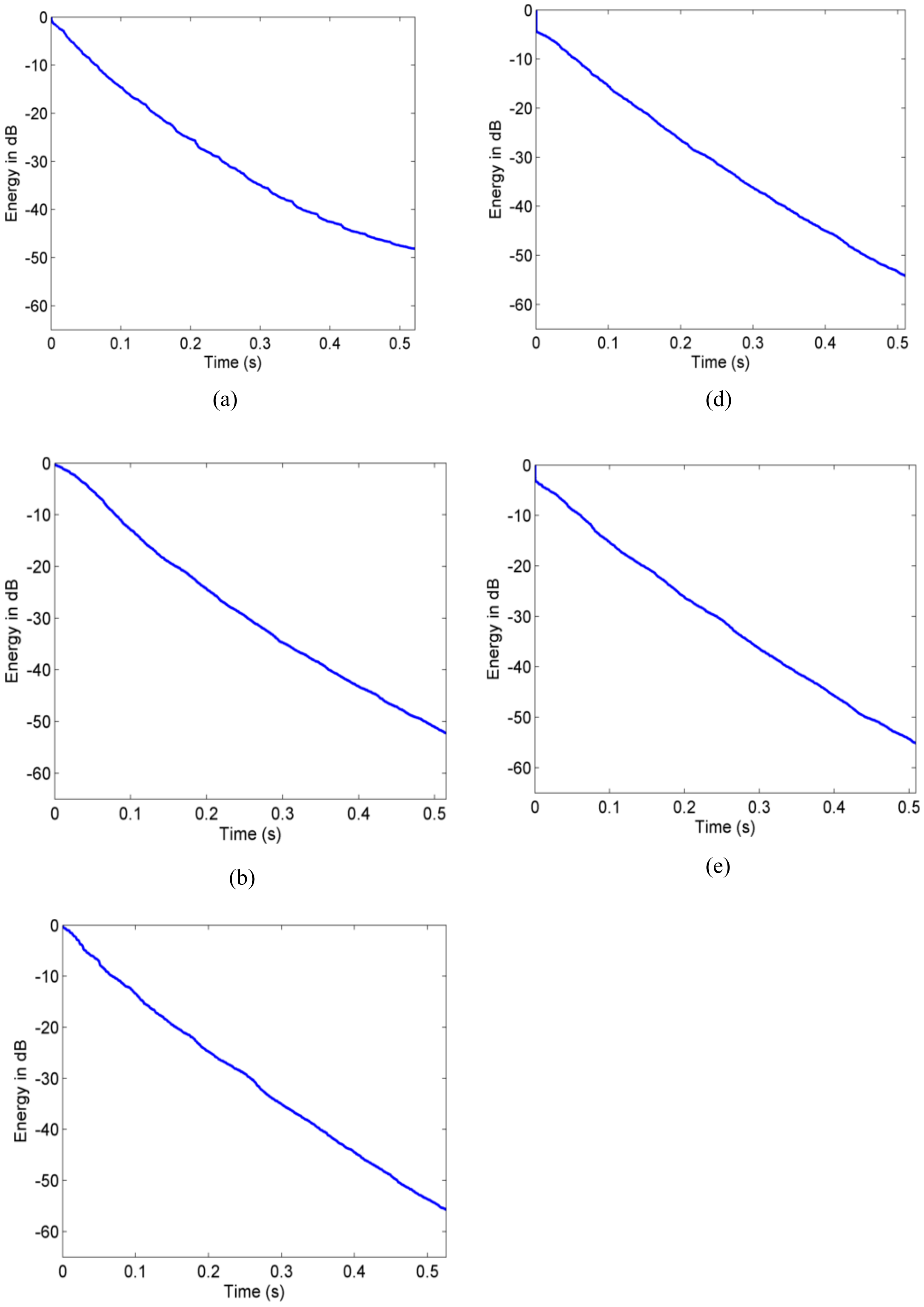
(d)

(b)

(e)

(c)

Fig. 4.3. Energy decay curves for (a) the original RIR with RT60 = 0.9 s, (b) the same RIR equalized by Algorithm 1, (c) the same RIR equalized by Algorithm 2, (d) the same RIR equalized by the algorithm of Wu and Wang [22], and (e) the same RIR equalized by the algorithm of Mosayyebpour *et al.* [26].



(a)

(d)

(b)

(e)

(c)

Fig. 4.4. Energy decay curves for (a) the original RIR with RT60 = 0.5 s, (b) the same RIR equalized by Algorithm 1, and (c) the same RIR equalized by Algorithm 2, (d) the same RIR equalized by the algorithm of Wu and Wang [22], and (e) the same RIR equalized by the algorithm of Mosayyebpour *et al.* [26].

Table 4.1. Estimated RT60 values for the original RIR and equalized RIRs by different methods for RT60 = 0.5 s and 0.9 s.

RT60 = 0.5 s					
Algorithm	Original RIR	Algorithm 1	Algorithm 2	Wu & Wang [22]	M. *et al.*[26]
Estimated RT60 (s)	0.52154	**0.4582**	0.50059	0.51151	0.5089
RT60 = 0.9 s					
Algorithm	Original RIR	Algorithm 1	Algorithm 2	Wu & Wang [22]	M. *et al.*[26]
Estimated RT60 (s)	0.91513	0.88444	**0.8612**	0.91617	0.96075

4.4. Normalized Segmental Signal to Reverberation Ratio

Fig. 4.5 shows the normalized segmental SRR values of the inverse-filtered speech signals by different algorithms for RIRs with three different reverberation times. The figure also depicts the scores of the reverberant speech for the purpose of comparison. By comparing the inverse-filtered signal by the proposed algorithms to that of Wu and Wang [22] and Mosayyebpour *et al.* [26] for three different reverberation times of $RT60 = 0.5\ s, 0.7\ s$ and $0.9\ s$, we see that in all the three cases, the inverse-filtering part of the proposed algorithms demonstrate a greater SRR score compared to their corresponding algorithms. In other words, Algorithm 1 shows a better dereverberation performance compared to that of Wu and Wang [22], both of which use kurtosis maxmization, and Algorithm 2 that uses skewness maximization outprforms the method of Mosayyebpour *et al.* [26], both of which use skewness maximization. The SRR level of the reverberant speech is found in the middle of the graph being much higher than the method of Wu and Wang [22] and, for the first two reverberation times, significantly higher than that of Algorithm 1. It is, in turn, lower than that of Mosayyebpour *et al.* [26] for the first two reverberation times and much lower than that of Algorithm 2 for all the three reverberation times. It is specifically interesting to note that the method of Mosayyebpour *et al.* [26] fails to maintain its performance for RIR with $RT60 = 0.9\,s$. Likewise, Algorithm 2, which similarly uses skewness maximization, experiences a significant drop in its SRR score for RIR with $RT60 = 0.9\,s$. This is while Algorithm 1, which is based on kurtosis maximization, does not experience such an incline. However, the SRR score of the inverse-filtering stage of Algorithm 2 is still significantly higher than that of the other algorithms even for

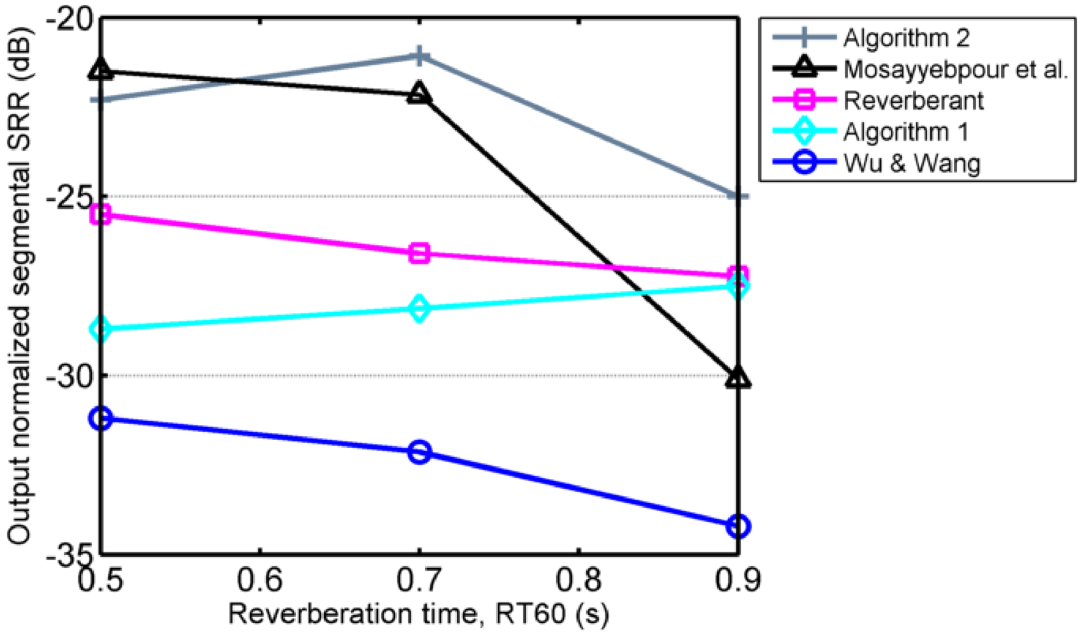


Fig. 4.5. Normalized segmental SRR values for reverberant speech and inverse-filtered speech signals by various algorithms in different reverberation times.

$RT60 = 0.9\,s$. Thus, it can be concluded that the inverse-filtering part of Algorithm 2 demonstrates the best performance among all the inverse-filtering methods compared here.

Fig. 4.6 depicts the normalized segmental SRR scores for the fully-processed speech signals by the two-stage algorithms, namely the algorithm of Wu and Wang depicts the SRR score of the reverberant speech signal. It can be easlisy seen that both Algorithms 1 and 2 outperform the method of Wu and Wang [22]. Algorithm 2, whose SRR score is well above that of the other algorithms, demonstrates the best performance with a substantial margin. Again, the reverberant speech, with a SRR score well below Algorithm 2, shows a higher SRR than that of Wu and Wang [22] but less than that of Algorithm 1 for the last two reverberation times. It should be noted that had we added the same second stage to the inverse-filtering algorithm of Mosayyebpour *et al.* [26], our Algorithm 2 would outperform the results of that

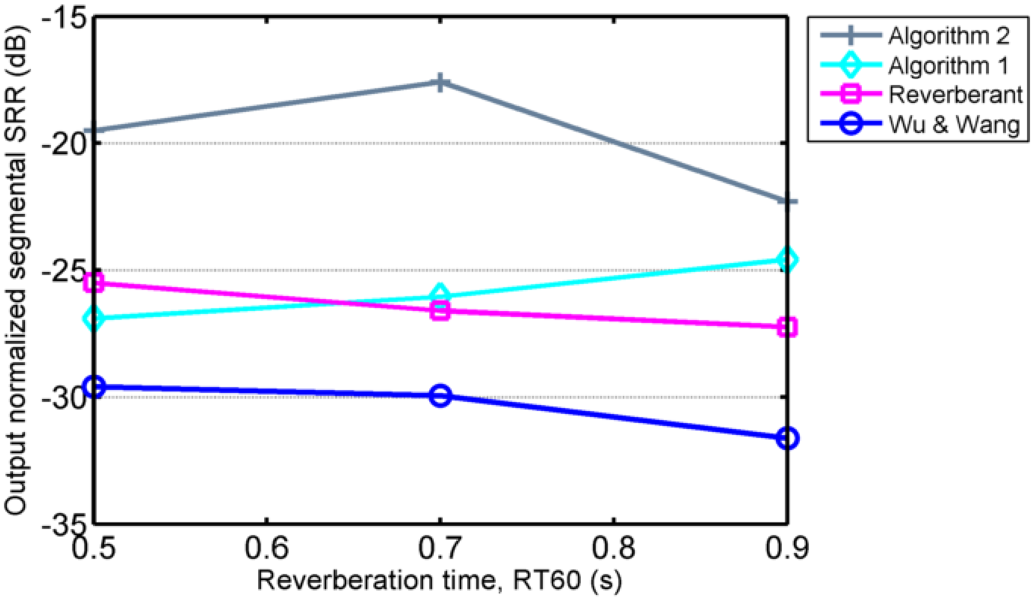


Fig. 4.6. Normalized segmental SRR values for reverberant speech and fully processed speech signals by various algorithms in different reverberation times.

algorithm as well, since it did so for the first stage (inverse-filtering).

4.5. Automatic Speech Recognition (ASR) and Perceptual Evaluation of Speech Quality (PESQ) Tests

While a review of the literature on the dereverberation evaluation casts doubt on the correlation of the results of ASR and PESQ measures with dereverberation as they are not directly developed for dereverberation assessment, they offer strong measures of the overall quality of the speech signal. Therefore, they can be employed along with other measures, which are known to have more correlation with dereverberation evaluation such as the normalized segmental signal to reverberation ratio.

The PESQ implementation is performed with the help of the MATLAB implementation associated with [77].[8] Both the narrowband and wideband implementation results are included.

The ASR measure[9], on the other hand, is a simulated automatic speech recognition test, which gives a confidence measure to assess the closeness of the text to the speech utterance associated with it. In other words, it is a test to simulate a subjective test performed on human listeners for which the result is shown as word error rates. The ASR simulation test is a solution to the subjective evaluation of word error rates that can be cumbersome and time-consuming.

The PESQ and ASR test results along with normalized SRR values of speech signals are given in Tables 4.2, 4.3 and 4.4. In addition, wideband PESQ scores for the same

[8] http://www.utdallas.edu/~loizou/speech/software.htm

[9] The ASR evaluation has been carried out based on the toolbox available at http://mirlab.org/jang/matlab/toolbox/asr/.

signals in three reverberation times are given in Table 4.5. These tables give the results for the first stage (inverse filtering) as well as for the complete algorithms in the case of the proposed algorithms and that of Wu and Wang [22] and Mosayyebpour *et al.* [26]. The reverberant, the inverse-filtered, and the fully-processed speech signals are indicated as ‘rev’, ‘inv’, and ‘proc’, respectively. As the method of Mosayyebpour *et al.* [26] is only an inverse-filtering algorithm and it does not include a second stage, and since they have addressed dereverberation as one of the applications of their inverse-filtering algorithm, the results of their algorithm are repeated in the column for the fully-processed signal. For each column, the best value is highlighted in bold.

Table 4.3. Summary results for the reverberant, the inverse-filtered and the fully-processed speech for RIR with reverberation time of 0.5 s.

Algorithm	Normalized segmental SRR (dB)			ASR (confidence measure)			PESQ score (NB)		
	rev	inv	proc	rev	inv	proc	rev	inv	proc
Algorithm 1	-25.5	-28.7	-26.9	78.94	**75.70**	71.50	2.20	2.04	1.81
Algorithm 2	-25.5	-22.3	**-19.5**	78.94	75.5	68.6	2.20	**2.14**	**1.85**
W. & W. [22]	-25.5	-31.2	-29.6	78.94	72.99	69.46	2.20	1.60	1.29
M. *et al.* [26]	-25.5	**-21.5**	-21.5	78.94	**75.70**	**75.70**	2.20	1.64	1.64

Table 4.2. Summary results for the reverberant, the inverse-filtered, and the fully-processed speech for RIR with reverberation time of 0.7 s.

Algorithm	Normalized segmental SRR (dB)			ASR (confidence measure)			PESQ score (NB)		
	rev	inv	proc	rev	inv	proc	rev	inv	proc
Algorithm 1	-26.6	-28.13	-26.1	75.02	**72.53**	68.56	2.06	1.95	1.77
Algorithm 2	-26.6	**-21.1**	**-17.6**	75.02	71.95	65.26	2.06	**2.06**	**1.83**
W. & W. [22]	-26.6	-32.1	-30.0	75.02	68.03	66.51	2.06	1.55	1.24
M. *et al.* [26]	-26.6	-22.2	-22.2	75.02	72.42	**72.42**	2.06	1.59	1.59

Table 4.4. Summary results for the reverberant, the inverse-filtered, and the fully-processed speech for RIR with reverberation time of 0.9 s.

Algorithm	Normalized segmental SRR (dB)			ASR (confidence measure)			PESQ score (NB)		
	rev	inv	proc	rev	inv	proc	rev	inv	proc
Algorithm 1	-27.23	-27.53	-24.58	68.48	**64.36**	56.87	1.97	**1.92**	1.69
Algorithm 2	-27.23	**-25**	**-22.3**	68.48	62.63	54.57	1.97	1.87	1.62
W. & W. [22]	-27.23	-34.21	-31.62	68.48	58.48	53.35	1.97	1.69	1.35
M. *et al.* [26]	-27.23	-30.1	-30.1	68.48	63.33	**63.33**	1.97	1.70	**1.70**

Table 4.5. Wideband PESQ scores for the reverberant, the inverse-filtered, and the fully-processed speech signals for RIRs with reverberation time values of 0.5, 0.7 and 0.9 s.

Algorithm	RT60=0.5 s			RT60=0.7 s			RT60=0.9 s		
	rev	inv	proc	rev	inv	proc	rev	inv	proc
Algorithm 1	1.44	1.33	1.26	1.34	1.28	1.23	1.28	**1.26**	**1.22**
Algorithm 2	1.44	**1.39**	**1.27**	1.34	**1.36**	**1.27**	1.28	**1.26**	1.21
W. & W. [22]	1.44	1.15	1.11	1.34	1.14	1.10	1.28	1.19	1.14
M. *et al.* [26]	1.44	1.16	1.16	1.34	1.15	1.15	1.28	1.19	1.19

As seen from Tables 4.2–4.4, the reverberant speech, in general, obtains greater score in PESQ and ASR tests compared to all the processed signals by all the algorithms in all the three reverberation times. This confirms the previously mentioned fact that these two measures are not correlated with dereverberation. However, it can be concluded that the proposed algorithms produce more intelligible speech compared to that produced by the existing algorithms, since both in the inverse-filtering and in spectral subtraction stages the PESQ values are higher than that for the two other algorithms.

On the other hand, as for the ASR test results, in the inverse-filtering stage, Algorithm 1 demonstrates superior results compared to all the other algorithms. Herein, it should be noted that, the results show that the spectral subtraction stage results in a reduced ASR score for the speech signal. Therefore, while repeating the same ASR score of the inverse-filtering algorithm of Mosayyebpour *et al.* [26] in the column for fully-processed speech signals, it is not surprising that this algorithm obtains the highest value. However, had we added the same second stage to this algorithm, the best ASR score would belong to Algorithm 1 in all the cases. In addition, in most cases, in terms of the ASR score, Algorithm 2 takes the third place after the method of Mosayyebpour *et al.* [26]. Considering the relatively high SRR score of this algorithm, one can conclude that there is a trade-off between suppressing reverberations and keeping the ASR score high. However, it should be noted that Algorithm 2 still outperforms the method of Wu and Wang in most cases [22].

Table 4.5 gives wideband PESQ scores for the same speech signals of the TIMIT database. It is noted that the wideband PESQ scores in general are more suitable for dereverberation, since they are not based on the assumption that the speech signal is restricted to the telephone band frequency spectrum. However, the table suggests that the wideband PESQ scores follow almost the same trend as the narrowband values do for different signals.

4.6. Spectrogram Improvement

Fig. 4.7 shows waveforms of a clean speech utterance and its reverberated version along with their corresponding spectrograms for RIR with a reverberation time of 0.9 s. In all the spectrograms presented in this work, since the voice activity level of the signal is important as it clearly affects the color map of the spectrograms, in order to have a proper comparison, all the signals are level-adjusted to zero activity level according to the ITU-T recommendation P.56[10]. The smearing effect of reverberation can be clearly seen both in the speech waveform and in the spectrogram, where the frequency pattern of the signal with respect to time is highly smeared.

Fig. 4.8 illustrates the waveforms and spectrograms of inverse-filtered speech signals by Algorithms 1 and 2 for the same speech utterance as in Fig. 4.7. Comparing the inverse-filtered signals from Fig. 4.8 to the clean and the reverberated signals in Fig. 4.7, it is noted that some reverberation effects have been removed. However, the spectrograms do not show a significant change at this stage.

[10] The MATLAB code for voice activity level adjustment was extracted from the toolbox VOICEBOX available at: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/activlev.html

Fig. 4.9 depicts the waveforms and spectrograms of inverse-filtered speech signals using the methods of Wu and Wang [22] and Mosayyebpour *et al.* [26] for the same speech utterance as in Fig. 4.7. Both from the spectrograms and the waveforms, it is seen that the inverse-filtered speech signals by these two methods are more smeared than the inverse-filtered speech signals by our proposed algorithms shown in Fig. 4.8. Between the two algorithms, however, Fig. 4.9 suggests that the one by the method of Mosayyebpour *et al.* [26] contains more smearing than the one by Wu and Wang [22] does. Moreover, by looking at Fig. 4.8 and Fig. 4.9, it can be concluded that, among the four, the inverse-filtered signal by the method of Mosayyebpour *et al.* [26] includes the largest amount of smearing by reverberation.

Fig. 4.10 depicts the waveforms and the spectrograms of the fully-processed speech signals by Algorithms 1 and 2 for the same speech utterance. By comparing these signals to the reverberant speech in Fig. 4.7, it can be clearly seen that the smearing effect is removed to a significant extent. Also, referring and comparing to the clean signal, between the two proposed algorithms, one may conclude that Algorithm 2 is more successful in dereverberation. However, although the difference between the processed signals by the two algorithms is clear, it is hard to pick one as a more successful algorithm only by looking at the waveforms and spectrograms.

The waveform and spectrogram of the fully-processed speech signal by the method of Wu and Wang [22] for the same speech utterance is depicted is Fig. 4. 11. Again, as compared to the reverberant speech in Fig. 4.7, the dereverberation effect is clear. On the other hand, by comparing this signal to those processed by our proposed algorithms, shown in Fig. 4. 10, one can conclude that the proposed algorithms leave less smearing. This smearing is detectable both in the waveform and in the spectrogram, where, in some regions with a high fluctuation of energy, which translates to sharp color changes in the spectrogram, the color contrast is recovered by the proposed algorithms, but it is lost in the processed signal by the method of Wu and Wang [22]. As a result, the overall pattern of the spectrogram is more preserved in the case of the proposed algorithms.

It is worth reminding that, since the method of Mosayyebpour *et al.* [26] shows inferior results in inverse filtering of the speech as compared to the proposed algorithms, although adding the same second stage to it would result in a better performance, the performance would be still inferior to that of the proposed algorithms.

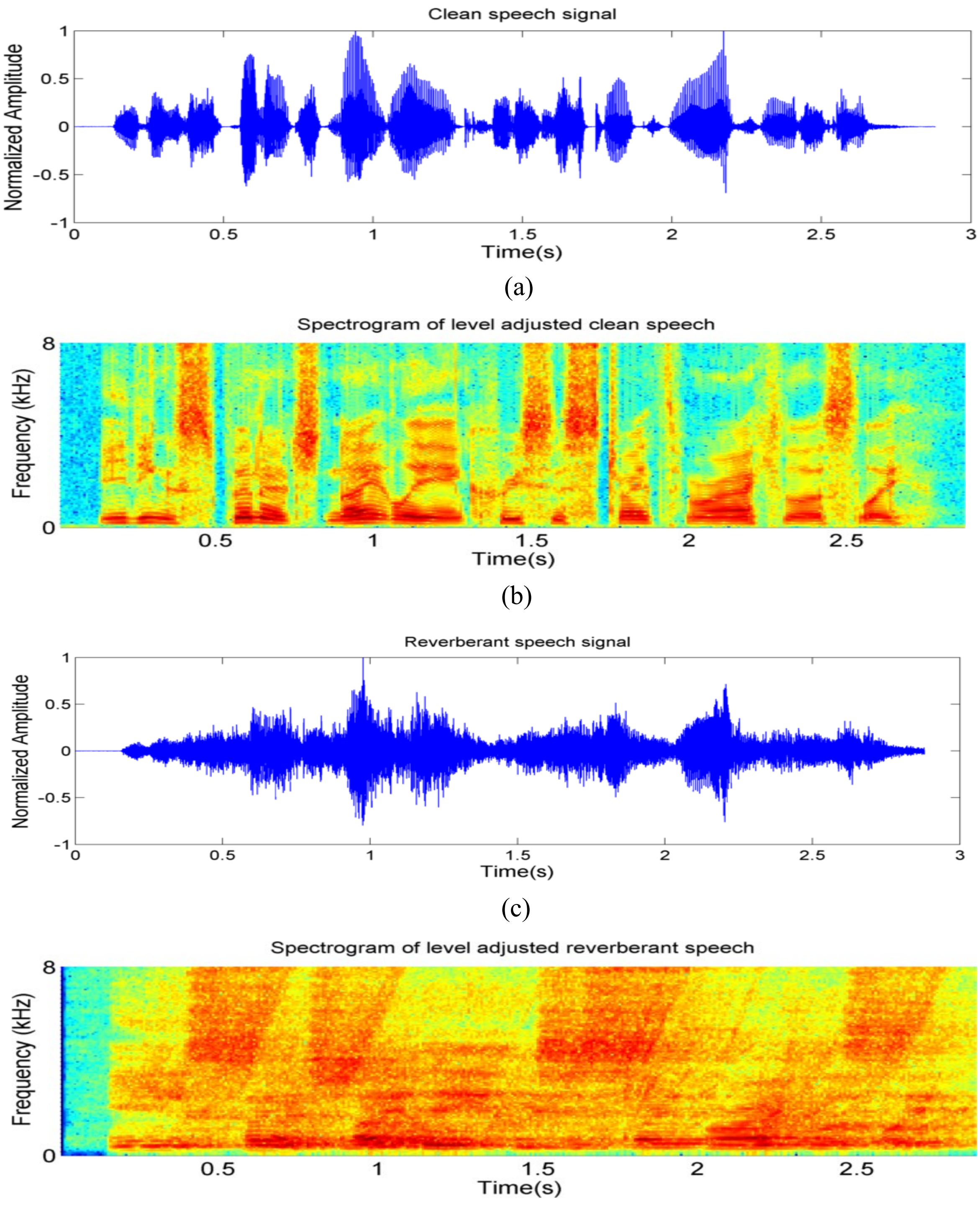
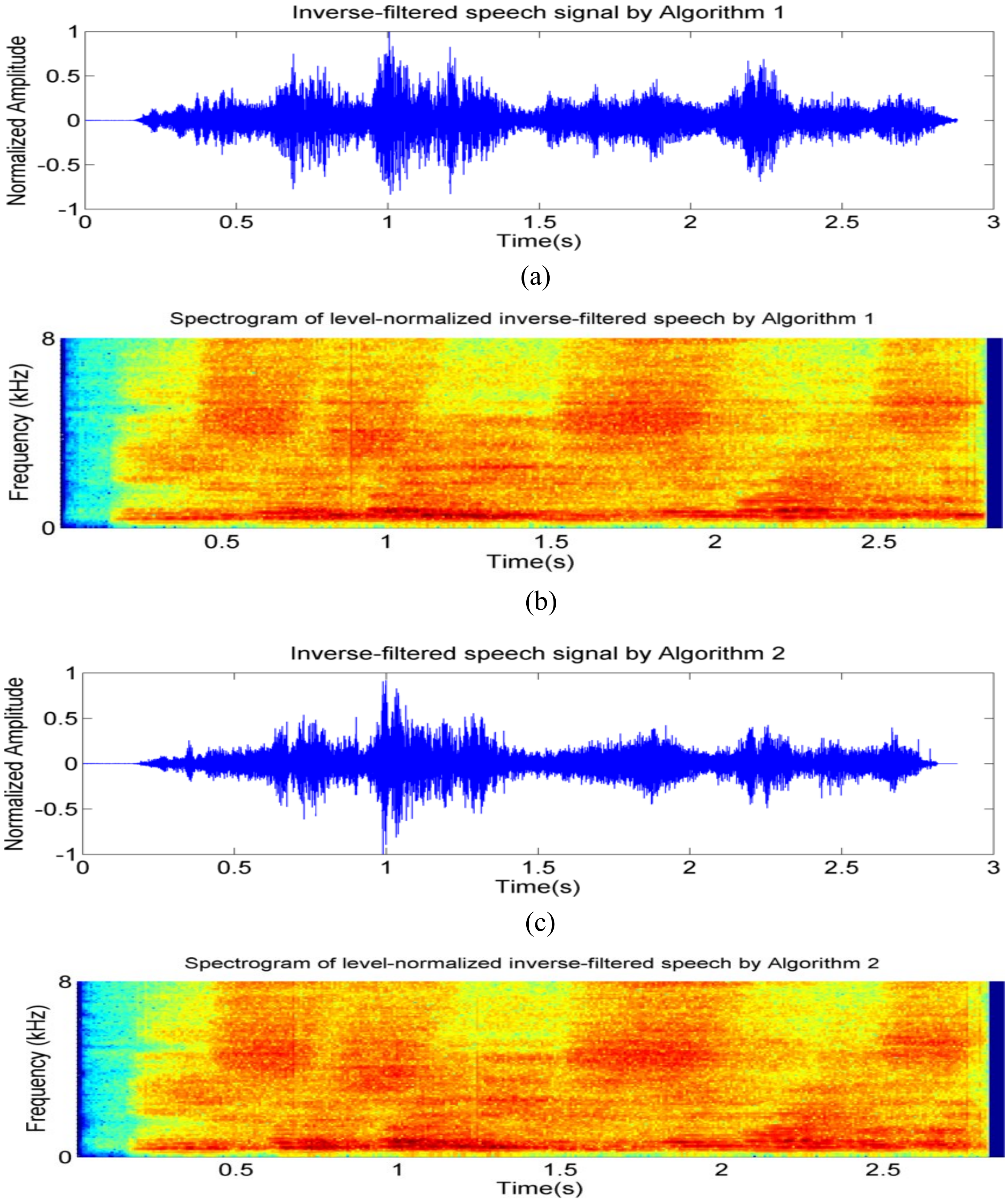


Fig. 4.7. A clean speech utterance from the TIMIT database and the associated reverberant speech signal along with the corresponding level-normalized spectrograms. The reverberant speech is produced by RIR with reverberation time of 0.9 s.



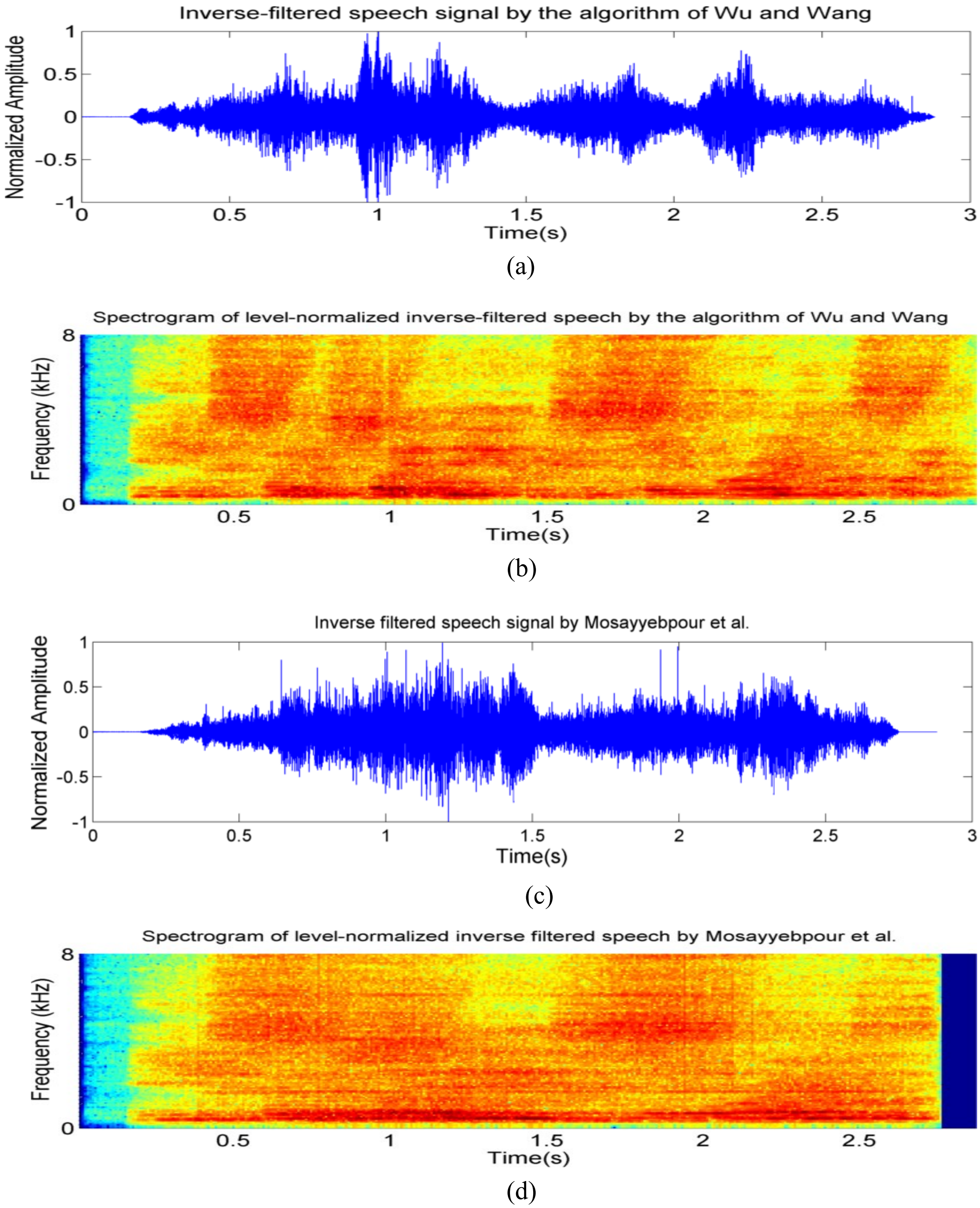
(a)

(b)

(c)

(d)

Fig. 4.8. The inverse-filtered speech signals by Algorithms 1 and 2 for the same speech utterance as in Fig. 4.7. along with the corresponding level-normalized spectrograms



Inverse-filtered speech signal by the algorithm of Wu and Wang

(a)

Spectrogram of level-normalized inverse-filtered speech by the algorithm of Wu and Wang

(b)

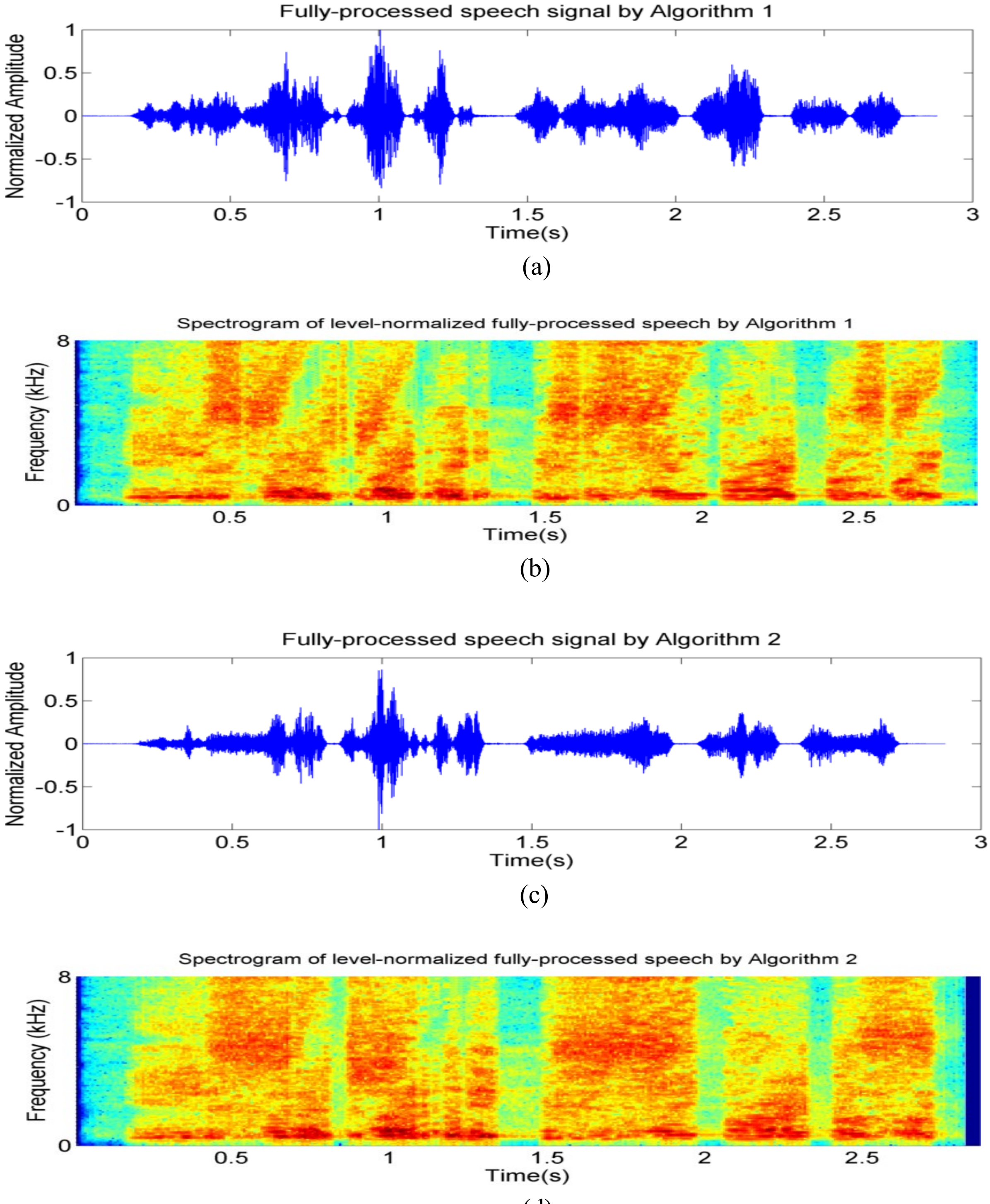
Inverse filtered speech signal by Mosayyebpour et al.

(c)

Spectrogram of level-normalized inverse filtered speech by Mosayyebpour et al.

(d)

Fig. 4.9. The inverse-filtered speech signals by the algorithm of Wu and Wang [22] and the algorithm of Mosayyebpour *et al.* [26] for the same speech utterance as in Fig. 4.7. along with the corresponding level-normalized spectrograms.



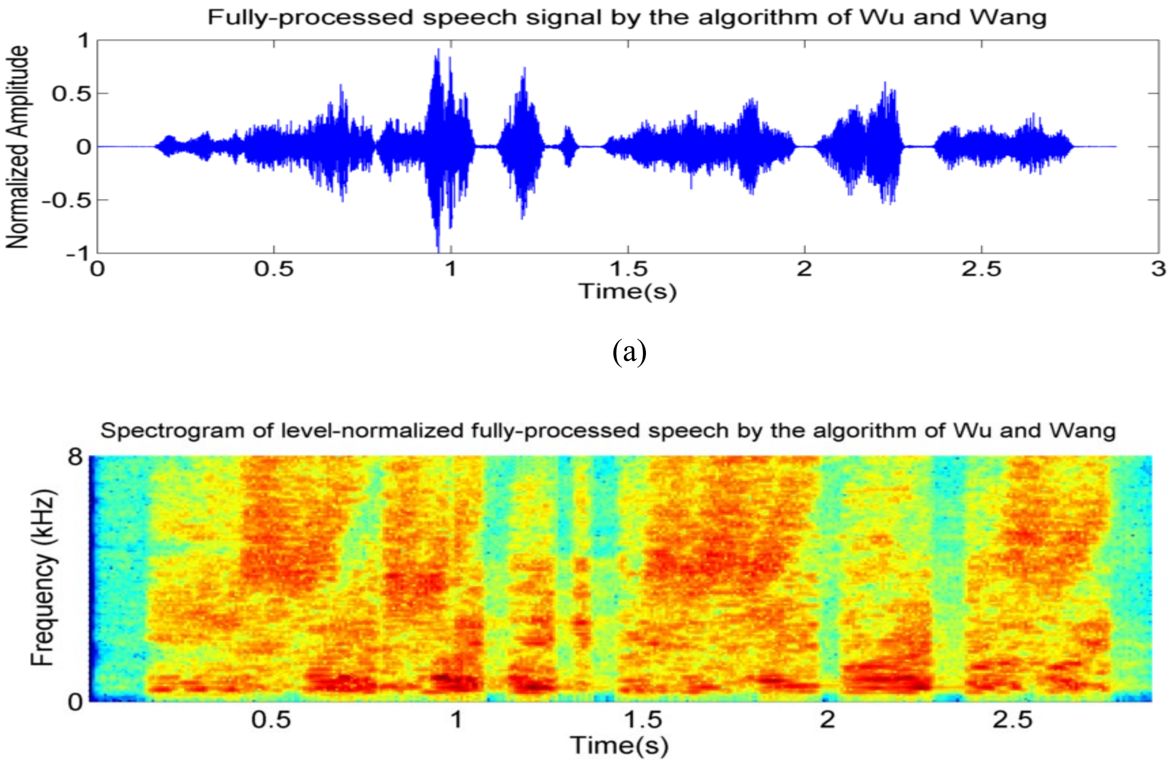
(a)

(b)

(c)

(d)

Fig. 4.10. The fully-processed speech signals by Algorithms 1 and 2 for the same speech utterance as in Fig. 4.7. along with the corresponding level-normalized spectrograms.



(a)

(b)

Fig. 4.11. The fully-processed speech signal by the algorithm of Wu and Wang [22] for the same speech utterance as in Fig. 4.7. along with the corresponding level-normalized spectrogram.

4.7. Robustness against Noise

In this section, the robustness of the proposed algorithms is investigated and compared to that of the other two algorithms in the presence of background noise. The noisy speech data are 30 speech utterances by a male speaker in the presence of a train noise in the background. The speech is mixed with the background noise at 4 different levels to provide SNR values from 0 dB (the noisiest) to 15 dB[11]. The normalized SRR results for the inverse-filtered speech signals and the fully-processed signals are obtained and

[11] The noisy speech data was extracted from NOIZEUS corpus available at http://www.utdallas.edu/~loizou/speech/noizeus/

compared for the four SNR levels. In this experiment, both the speech signals and the RIRs are sampled at $f_s = 8\ kHz$.

Fig. 4.12 contains the normalized SRR results of the inverse-filtered speech signals for all the four inverse-filtering algorithms for four input SNR levels for three different reverberation times. For comparison, the normalized SRR level of the reverberant signal is also depicted. As can be seen, for almost all the four SNR levels and for all RIRs, the inverse-filtering of the proposed algorithms offers better SRR results as compared to the other two algorithms. Among the two proposed algorithms, inverse-filtering by Algorithm 2 shows a significantly better performance. As compared to the reverberant signal, only for RIR with $RT60$ of 0.9 s the inverse-filtered speech by Algorithm 2 improves the SRR in noisy background situation in three out of the four SNR levels. In all other cases the reverberant signal has the best SRR score.

In Fig. 4.13, the normalized SRR results of the fully-processed speech signals by the three two-stage algorithms for four input SNR levels for three reverberation times are depicted. Again, the normalized SRR level of the reverberant speech signal is also shown. It can be seen that the best SRR score in all cases belongs to Algorithm 2. The reverberant signal, Algorithm 1, and the method of Wu and Wang [22] take up the next positions, respectively.

The modified Bark spectral distortion (MBSD) scores for the inverse-filtering stage of the algorithms for the four SNR levels for three reverberation times are shown in Fig. 4.14. The figure indicates that in almost all the cases, among the four inverse-filtering algorithms, Algorithm 2 suffers from the least distortion. In general, Algorithm 1, the methods of Mosayyebpour *et al.* [26] and Wu and Wang [22] occupy the next positions, respectively. In addition, especially for the two higher SNR levels, the reverberant signal shows the least amount of distortion as compared to the inverse-filtered signals using the various algorithms.

Fig. 4.15 depicts the MBSD score results for the fully-processed speech signals using the different algorithms in the four input SNR levels for three reverberation times. The normalized SRR level of the reverberant speech is also included in the graphs. The figure indicates that for input SNR levels of 10 dB and less, in most cases, Algorithm 2, Algorithm 1, the method of Wu and Wang [22] and the reverberant signal in that order suffer from least distortion. However, for input SNR value of 15 dB, and especially for RIR with reverberation time of 0.9 s, Algorithm 2 loses the first place while Algorithm 1 shows the best performance among the three algorithms.

Finally, it is to be noted that many of the algorithms proposed in the literature for speech dereverberation provide stable performance only for input speech signals with higher SNR levels. For instance, Kinoshita *et al.* [27] have reported a stable performance of their algorithms for SNR levels higher than 20 dB, while the algorithms proposed in this thesis show stable results for SNR levels as low as 0 dB.

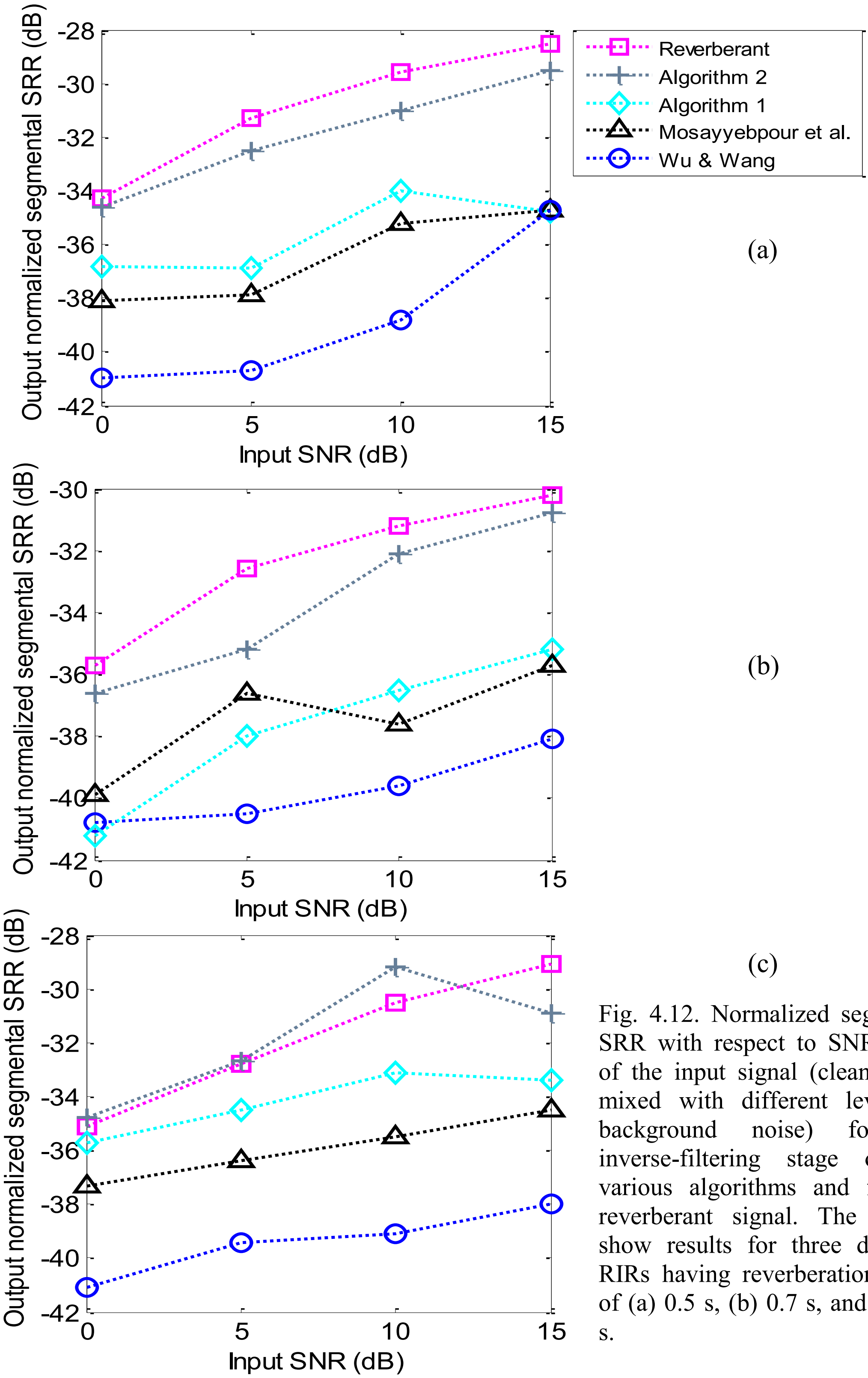


Fig. 4.12. Normalized segmental SRR with respect to SNR value of the input signal (clean signal mixed with different levels of background noise) for the inverse-filtering stage of the various algorithms and for the reverberant signal. The graphs show results for three different RIRs having reverberation times of (a) 0.5 s, (b) 0.7 s, and (c) 0.9 s.

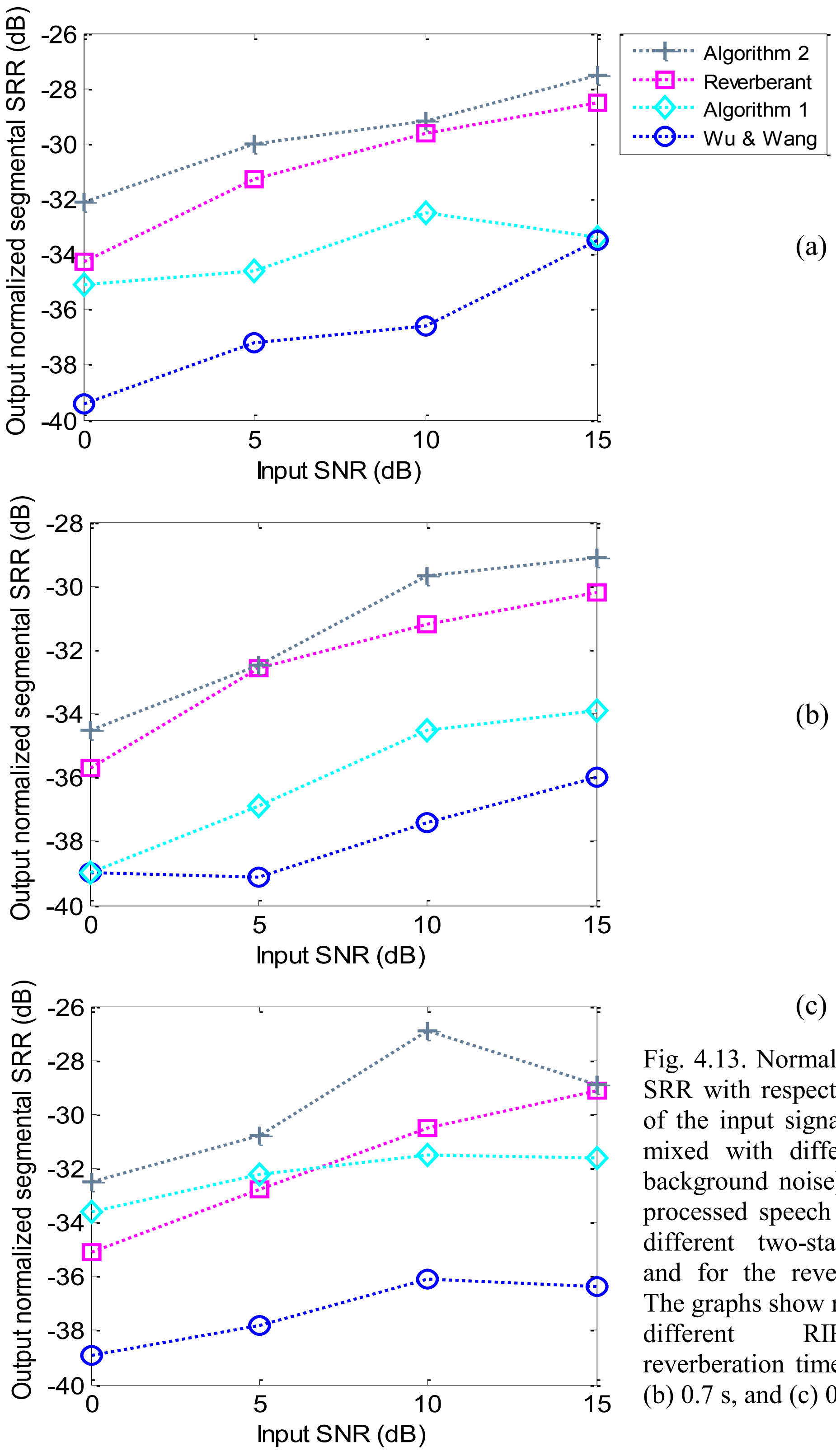
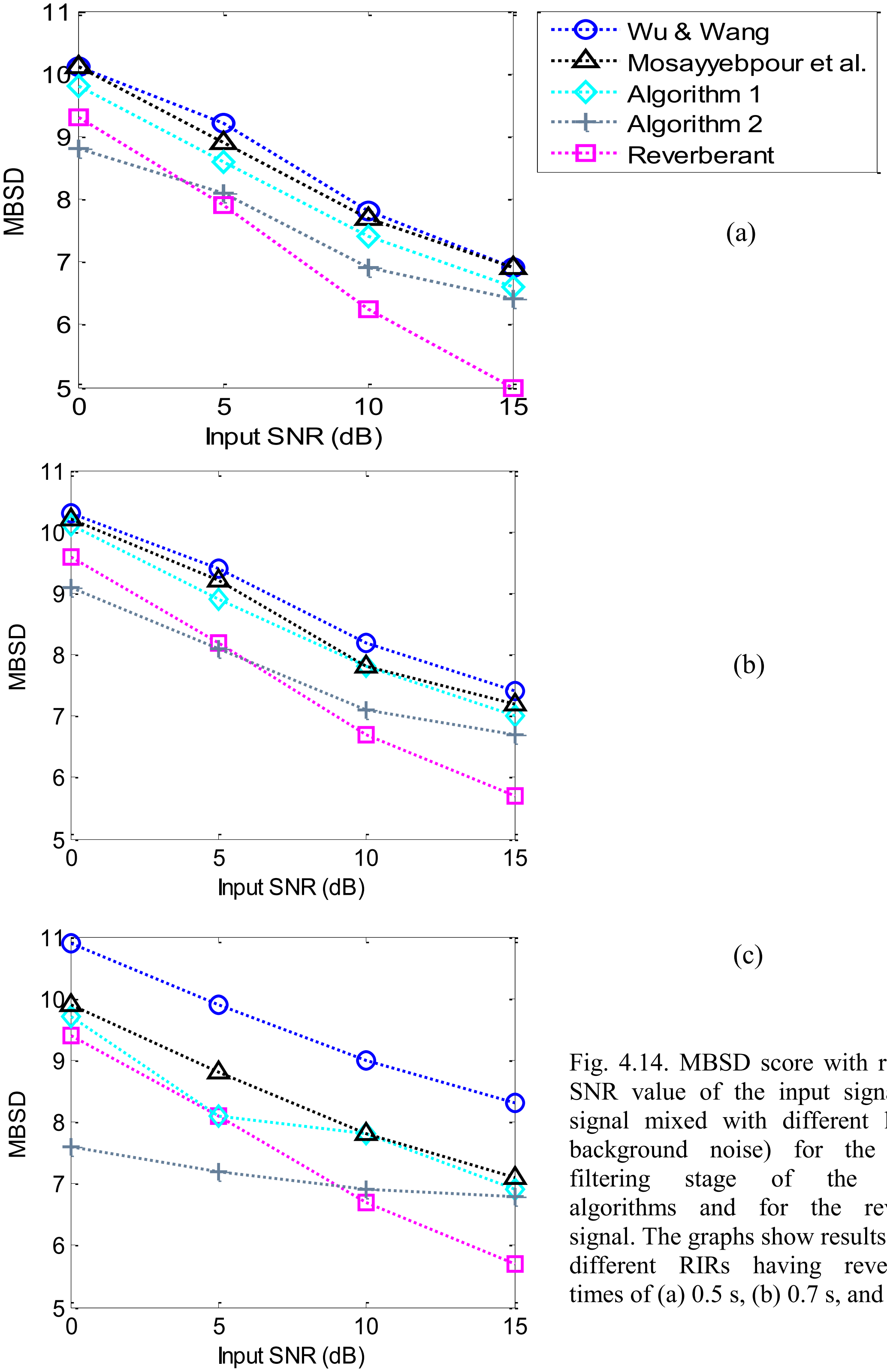


Fig. 4.13. Normalized segmental SRR with respect to SNR value of the input signal (clean signal mixed with different levels of background noise) for the fully-processed speech signals by the different two-stage algorithms and for the reverberant signal. The graphs show results for three different RIRs having reverberation times of (a) 0.5 s, (b) 0.7 s, and (c) 0.9 s.



(a)

(b)

(c)

Fig. 4.14. MBSD score with respect to SNR value of the input signal (clean signal mixed with different levels of background noise) for the inverse-filtering stage of the different algorithms and for the reverberant signal. The graphs show results for three different RIRs having reverberation times of (a) 0.5 s, (b) 0.7 s, and (c) 0.9 s.

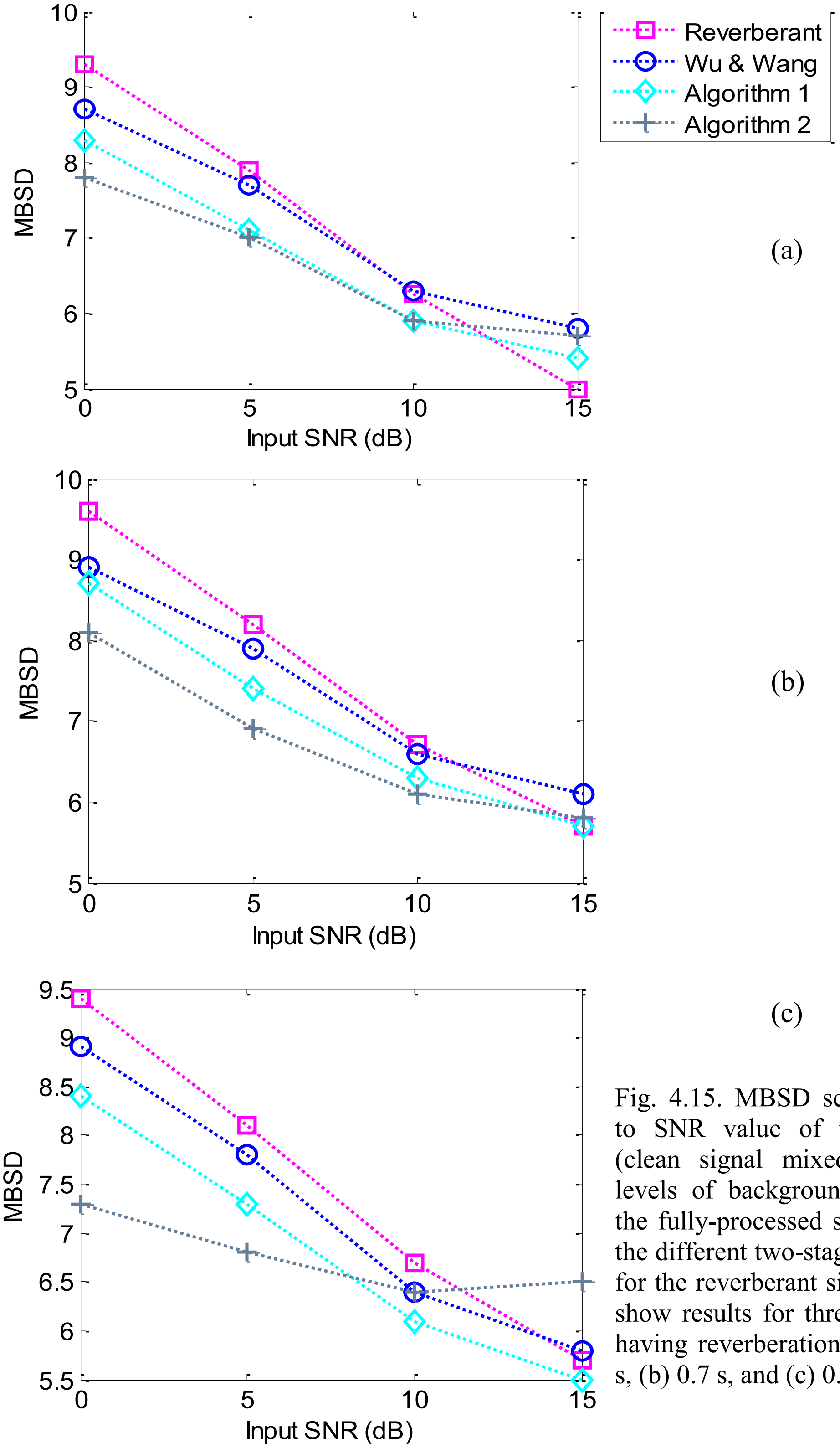


Fig. 4.15. MBSD score with respect to SNR value of the input signal (clean signal mixed with different levels of background noise) for the the fully-processed speech signals by the different two-stage algorithms and for the reverberant signal. The graphs show results for three different RIRs having reverberation times of (a) 0.5 s, (b) 0.7 s, and (c) 0.9 s

4.8. Summary

In this chapter, the experimental setup and the parameter settings of the implementation of the algorithms were first described. Then results concerning the performance of the proposed algorithms were obtained using some of the most relevant and frequently used qualitative and quantitative measures and compared to that of two of the most well-known algorithms. It has been shown that the equalized RIRs by the inverse-filtering stage of the proposed algorithms result in an overall less reverberation time. In particular, some of the mid to late impulses have been more successfully removed; at the same time, the overall structure of the RIR remained more intact which results in a more natural sounding speech. Also, the normalized segmental SRR of the proposed algorithms in general is higher compared to that of the existing algorithms. This is true both when only the inverse-filtering is considered and when the full algorithms are examined. In order to compare the overall perception quality and the overall automatic speech recognition performance of the algorithms, PESQ (narrowband and wideband) and ASR simulation results have also been obtained. In most cases, the scores were in favor of one or the other of the two proposed algorithms. The waveforms and the spectrograms of the clean, the reverberant, the inverse-filtered and the fully-processed signals by various algorithms have also been obtained. A close examination of some of the important regions of these waveforms and the spectrograms (mainly the regions with a higher energy contrast), has shown the superiority of the proposed algorithms in suppressing the reverberant components and recovering the clean signal. Finally, the relative robustness of the proposed algorithms against background noise was investigated by measuring the normalized segmental SRR and the modified Bark spectral distortion scores in four different input SNR levels; the results have confirmed that the proposed algorithms are able to maintain the dereverberation efficiency even for highly contaminated speech signals of SNRs close to zero.

Finally, it should be emphasized that, since the method of Mosayyebpour *et al.* [26] consists of the inverse-filtering only, it has not been included when comparing the fully-processed speech signals by the two-stage algorithms. However, since the results indicate that the inverse-filtering stage of our proposed algorithms outperforms that of Mosayyebpour *et al.* [26], after adding the same second stage to that algorithm, the proposed algorithms still would outperform the resulting two-stage method.

Chapter 5

Conclusion and Future Work

5.1. Concluding Remarks

This thesis has been concerned with the problem of single-microphone dereverberation. The main objective of this work has been to propose new algorithms that are more efficient than the existing ones in suppressing both short and long reverberation components for short and long room impulse responses (RIRs). Based on a critical examination of the performance of some of the major previous works, two new two-stage algorithms have been proposed in this thesis. The proposed algorithms have been shown to meet the above-mentioned goal and to be more robust against background noise.

The first stage of the proposed algorithms, inverse filtering, consists of pre-whitening followed by a delayed long-term LP filtering, whose kurtosis or skewness of the LP-residuals is maximized to control the weight updates of the inverse filter. Due to the convergence problem in a time domain implementation, the kurtosis or skewness maximization and the inverse-filtering have been carried out in the frequency domain. The short-term LP for pre-whitening and the delayed long-term LP together make up the first stage of the proposed algorithms. In the second stage, to further improve the dereverberation performance, a nonlinear spectral subtraction scheme has been employed.

The two proposed algorithms have been referred to as Algorithm 1 or Algorithm 2 depending on whether kurtosis or skewness of the LP-residual is maximized to control the weight updates of the inverse filter in the first stage.

Algorithm 2 utilizes less delay than Algorithm 1 does and it is more sensitive to the adaptation of the inverse-filter length to the reverberation time. In view of this, the optimal inverse-filter length for each reverberation time has been obtained empirically for Algorithm 2.

It has been shown that the proposed algorithms outperform some of the existing major dereverberation algorithms in terms of a number of qualitative and quantitative measures, such as equalized impulse responses and their energy decay curves and normalized segmental signal-to-reverberation ratio. Perceptual evaluation of speech quality and automatic speech recognition simulation results have also been included to compare the overall quality of the processed signals. Finally, an investigation has been carried out to demonstrate the robustness of the proposed algorithms against background noise.

It is concluded that the proposed algorithms are more efficient in dereverberation of short and long reflections for RIRs with different reverberation times. In addition, they are more robust against the background noise. Moreover, among the two proposed algorithms, Algorithm 2, the one using skewness maximization, is more successful in dereverberating the speech signal. This has been particularly inferred from the comparison of the signal-to-reverberation ratio results of the proposed algorithms, in which there is a significant difference between the two algorithms. In most of the other aspects of the comparison, the difference between the two proposed algorithms is not as significant.

5.2. Scope for Future Work

The complexity of the proposed algorithms is relatively high. The software implementation of the proposed algorithms results in a processing time that is 2 to 3 times higher than that using the method of Wu and Wang [22]. In MATLAB this means a 45-second processing time for a 3 to 4 second speech segment. As pointed out at the end of Section 3.2.1, the higher execution time of the proposed technique is firstly due to the long-term linear prediction in DLLP and secondly due to having prewhitening as another phase of linear prediction. Clearly, implementing the algorithms in a high-level programing language, such as C/C++, would drastically reduce the processing time of the algorithms possibly to an extent that might make them suitable even for real-time applications. This claim, of course, needs to be confirmed only after implementing the proposed algorithms in a high-level language. Even though the processing time of the proposed algorithms in MATLAB is acceptable for non-real time applications, a hardware implementation of the algorithms could be another task that could be undertaken in future after assessing the algorithms by implementing them in a high-level language such as C/C++. Only then, the final performance of the algorithms can be truly assessed. The possible hardware implementation can then be used, for example, as an integrated component of a device, which works with speech commands or other devices used in speech communication where the quality of the speech signal for recognition is important.

References

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation.* Berlin, Germany: Springer, 2010.

[2] R. H. Bolt and A. D. MacDonald, “Theory of speech masking by reverberation,” *J. Acoust. Soc. Am.*, vol. 21, no. 6, pp. 577–580, 1949.

[3] H. Haas, “The influence of a single echo on the audibility of speech,” *J. Audio Eng. Soc.*, vol. 20, pp. 145–159, 1972.

[4] E.A.P Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” Ph.D. dissertation, T. U. Eindhoven, 2007. Available: http://alexandria.tue.nl/extra2/200710970.pdf.

[5] J.S. Garofolo, “Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database,” National Institute of Standards and Technology, Gaithersburg, Maryland, Tech. Rep., Dec. 1988.

[6] X. Huang, A. Acero, and H. Hon, *Spoken language processing. a guide to theory, algorithm and system development*. Prentice-Hall, 2001.

[7] D. Bees, M. Blostein, and P. Kabal, “Reverberant speech enhancement using cepstral processing,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP),* 1991, pp. 977–980.

[8] C. Avendano and H. Hermansky, “Study on the dereverberation of speech based on temporal envelope filtering,” in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 1996, pp. 889–892.

[9] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, “A method based on the MTF concept for dereverberating the power envelope from the reverberant signal,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2003, pp. 840–843.

[10] T. Nakatani and M. Miyoshi, “Blind dereverberation of single channel speech signal based on harmonic structure,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2003, pp. 92–95.

[11] T. Nakatani, K. Kinoshita, and M. Miyoshi, “Harmonicity-Based Blind Dereverberation for Single-Channel Speech Signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol.15, no.1, pp.80-95, Jan. 2007.

[12] E. A. P. Habets, S. Gannot, and I. Cohen. “Late reverberant spectral variance estimation based on a statistical model,” *IEEE Signal Process. Lett.*, vol. 9, pp. 770–773, 2009.

[13] K. Lebart and J. Boucher, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[14] J. S. Erkelens and R. Heusdens, “Single-microphone late-reverberation suppression in noisy speech by exploiting long-term correlation in the DFT domain,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2009, pp. 3997–4000.

[15] E. A. P. Habets, S. Gannot, and I. Cohen, “Speech dereverberation using backward estimation of the late reverberant spectral variance,” in *Proc. IEEE Conf. Elect. Electron. Engineers in Israel*, Dec. 2008, pp. 384–388.

[16] K. Kinoshita, T. Nakatani, and M. Miyoshi, “Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2006, pp. 817–820.

[17] H. W. Löllmann and P. Vary, “Low delay noise reduction and dereverberation for hearing aids,” *EURASIP J. Appl. Signal Process.*, vol. 1, 2009.

[18] B. Yegnanarayana, “Enhancement of reverberant speech using LP residual,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 1998, vol. 1, pp. 405–408.

[19] B. Yegnanarayana and P.S. Murthy, “Enhancement of reverberant speech using LP residual signal,” *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, 2000.

[20] P. Krishnamoorthy and S. R. M. Prasanna, “Reverberant speech enhancement by temporal and spectral processing,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 137–148, Feb. 2009.

[21] B.W. Gillespie, H. Malvar, and D. Florêncio, “Speech dereverberation via maximum kurtosis subband adaptive filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2001, vol. 6, pp. 3701–3074.

[22] M. Wu and D. L. Wang, “A two-stage algorithm for one microphone reverberant speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.

[23] D. Fee, C. Cowan, S. Bilbao, and I. Ozcelik, “Predictive deconvolution and kurtosis maximization for speech dereverberation,” in *Proc. European Signal Process. Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006.

[24] S. Mosayyebpour, M. Esmaeili, T. A. Gulliver, “Single-Microphone Early and Late Reverberation Suppression in Noisy Speech,” *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 21, no. 2, pp. 322–335, Feb. 2013.

[25] E. A. P. Habets, N. Gaubitch, and P. A. Naylor, “Temporal selective dereverberation of noisy speech using one microphone,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2008, pp. 4577–4580.

[26] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "Single-Microphone LP Residual Skewness-Based Inverse Filtering of the Room Impulse Response," *IEEE Trans. Audio, Speech, Lang. Process.*, vol.20, no.5, pp.1617-1632, July 2012.

[27] 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. 4, May 2009.

[28] H. Kuttruff, *Room acoustics*, 4th ed. Taylor & Francis, 2000.

[29] J. Benesty, M. M. Sondhi, and Y. A. Huang, *Springer Handbook of Speech Processing*. Springer, 2008.

[30] W. C. Sabine, *Collected papers on acoustics*. Dover Publications, 1964.

[31] H. Kuttruff, *Room acoustics*. 4 ed. Taylor & Francis, 2000.

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

[33] 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, 2003.

[34] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, 1979.

[35] W. C. Sabine, *Collected papers on acoustics.* Peninsula Publishing, 1993.

[36] J. A. Moorer, "About this reverberation business," *Comput. Music J.*, no. 2, pp. 13–28, 1979.

[37] J. D. Polack, "La transmission de I'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, 1998.

[38] K. Kokkinakis and P. C. Loizou, "Evaluation of Objective Measures for Quality Assessment of Reverberant Speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2011, pp. 2420-2423.

[39] "Methods for subjective determination of transmission quality," Recommendation P.800, International Telecommunications Union (ITU-T), Feb. 1996.

[40] D. Picovici and A.E. Mahdi, "Towards non-intrusive speech quality assessment for modern telecommunications," in *First Joint IEI/IEE Symp. Telecom Systems Research*, Nov. 2001.

[41] 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, Nov 2006.

[42] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, *Objective Measures of Speech Quality*, O. Englewood Cliffs, New Jersey: Prentice-Hall, 1988.

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

[44] P.A. Naylor and N.D. Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, 2005.

[45] H. Wang and F. Itakura, "An implementation of multi-microphone dereverberation approach as a preprocessor to the word recognition system," *J. Acoust. Soc. Jap.*, vol. 13, no. 5, pp. 285–293, 1992.

[46] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: MacMillan, 1993.

[47] W. Yang, M. Dixon, and R. Yantorno, "A modified bark spectral distortion measure which uses noise masking threshold," in *IEEE Speech Coding Workshop*, Pocono Manor, 1997, pp. 55–56.

[48] W. Yang, "Enhanced Modified Bark Spectral Distortion (EMBSD): an objective quality measure based on audible distortion and cognition model," Ph.D. dissertation, Temple University, May 1999.

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

[50] ITU-T recommendation P. 862.2. Available: http://www.itu.int/rec/T-REC-P.862.2-00711-I/en.

[51] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, M. Brookes and P. A. Naylor, "C-Qual - a validation of PESQ using degradations encountered in forensic and law enforcement audio," in *Proc. AES 39TH Int. Conf.*, Hillerød, Denmark, June 2010.

[52] K. Kokkinakis and P. C. Loizou, "Evaluation of Objective Measures for Quality Assessment of Reverberant Speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2011, pp. 2420-2423.

[53] ITU-T recommendation P. 863. Available: http://www.itu.int/rec/T-REC-P.863/

[54] P.A. Naylor, N.D. Gaubitch and E.A.P Habets, "Signal-based performance evaluation of dereverberation algorithms," *J. Elect. Comput. Eng. (formerly Hindawi Research Letters in Signal Processing)*, 2010.

[55] J. Hardwick, C.D. Yoo, and J.S. Lim, "Speech enhancement using the dual excitation speech model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 1993, pp. 367–370.

[56] C.D. Yoo, "Speech enhancement based on the generalized dual excitation model with adaptive analysis window," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 1995, vol. 1, pp. 832–835.

[57] M.S. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 1998, vol. 6, pp. 3613–3616.

[58] M. Brandstein and S. Griebel, "Explicit Speech Modeling for Microphone Array Applications," in *Microphone arrays: signal processing techniques and applications*, M. Brandstein and D. Ward, Springer, 2001, ch. 7, pp. 133–153.

[59] H. Attias and L. Deng, "Speech Denoising and Dereverberation Using Probabilistic Models," *Advances in Neural Inform. Process. Syst.*, vol. 13, pp. 758–764, 2001.

[60] T. Nakatani, B.H. Juang, K. Kinoshita, and M. Miyoshi, "Speech dereverberation based on probabilistic models of source and room acoustics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2006, vol. 1, pp. 821-824.

[61] J.B. Allen, "Synthesis of pure speech from a reverberant signal," U.S. Patent 3786188, 1974.

[62] S. Griebel and M. Brandstein, "Wavfelet tranform extrama clustering for multi-hannel speech deverberation," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 1999.

[63] S. Griebel and M. Brandstein, "Microphone array speech dereverberation using coarse channel modeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Nov. 2001, vol. 1, pp. 201–204.

[64] M. Tonelli, N. Mitianoudis, and M.E. Davies, "Maximum Likelihood approach to blind audio de-reverberation," in *Proc. 7th Int. Conf. Digital Audio Effects (DAFx)*, Naples, Italy, Oct. 2004, pp. 1–6.

[65] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-Interscience, 2001.

[66] M. Tonelli, M.G. Jafari, and M.E. Davies, "A multi-channel Maximum Likelihood approach to de-reverberation," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Florence, Italy, Sept. 2006.

[67] B. Yegnanarayana, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2002, vol. 1, pp. 541–544.

[68] N.D. Gaubitch, P.A. Naylor, and D.B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 809–812.

[69] N.D. Gaubitch, P.A. Naylor, and D. Ward, “On the use of linear prediction for dereverberation of speech,” in *Proc. of the Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Kyoto, Japan, 2003, pp. 99–102.

[70] N.D. Gaubitch, D.B. Ward, and P.A. Naylor, “Statistical analysis of the autoregressive modeling of reverberant speech,” *J. Acoust. Soc. Am.*, vol. 120, pp. 4031–4039, Dec. 2006.

[71] D. Gesbert and P. Duhamel, “Robust blind identification and equalization based on multi-step predictors,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 1997, vol. 26, no. 5, pp. 3621–3624.

[72] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, Second Edition*, Revised and Expanded.* New York: Marcel Dekker, 2001.

[73] D. A. Harville, *Matrix Algebra from a Statistician’s Perspective.* New York: Springer, 1997.

[74] O. Tanrikulu and A.G. Constantinides, “Least-mean kurtosis: a novel higher-order statistics based adaptive filtering algorithm,” *Electronics Lett.*, vol. 30, no. 3, pp. 189–190, Feb., 1994.

[75] S. Haykin, *Adaptive Filter Theory, 4th ed.* Upper Saddle River, N.J.: Prentice-Hall, 2002.

[76] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[77] Y. Hu, and P. Loizou, “Evaluation of objective quality measures for speech enhancement *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229-238, Jan 2008.