BI-MICROPHONE SPEECH ENHANCEMENT
RESTAURATION DE LA PAROLE À DEUX MICROPHONES

Master's Thesis
Speciality: Electrical Engineering

Jonathan FILLION-DENEAULT

Sherbrooke (Québec) Canada  August 2006
NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

AVIS:
L’auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.
RÉSUMÉ

La réduction de bruit acoustique est une opération qui cherche à atténuer le bruit présent dans un enregistrement de voix bruité. Que ce soit pour la téléphonie mobile, pour les prothèses auditives électroniques ou pour les systèmes de commande vocale d'automobiles, le problème de débruitage est de plus en plus préoccupant, surtout si on considère l'augmentation exponentielle des utilisateurs de ces technologies. Plusieurs solutions existent déjà à l'heure actuelle pour palier au problème de débruitage et augmentent en général l'intelligibilité et la qualité du signal de parole traité; les systèmes sont généralement classés en deux catégories, soit ceux à un seul microphone et ceux à plusieurs microphones. Les systèmes à un seul microphone exploitent principalement les propriétés statistiques différentes de la parole et du bruit afin de séparer ces derniers alors que les systèmes à plusieurs microphones exploitent plutôt la diversité spatiale du locuteur et des sources de bruit, ce qui requiert le plus souvent un positionnement et une calibration délicate des microphones. Cette recherche présente un système à deux microphones qui vise à combler le vide entre ces deux extrêmes. La cohérence de la parole entre les deux microphones est exploitée pour construire un post-filtre servant à compléter les méthodes spectrales à un seul microphone utilisant les meilleures lois de suppression et d’estimation du bruit. Le système proposé est ainsi simple et robuste, imposant à la fois de faibles restrictions quant à l'emplacement et la directivité des microphones utilisés, tout en facilitant une intégration rapide sur les appareils portables existants. Des tests objectifs et subjectifs attestent de la qualité supérieure des échantillons traités par le système proposé par rapport aux meilleures techniques à un seul microphone.
ABSTRACT

Speech enhancement is an operation that aims to reduce acoustic noise picked up by noisy recordings of speech. It is essential to various systems such as mobile communications, hearing aids and hands free car telephony where the speaker's voice is often sampled in adverse acoustical conditions. Various solutions and implementations already exist to enhance the intelligibility and quality of speech in noisy environments and can generally be classified in two different groups; single microphone and multi-microphone methods. Single microphone techniques rely on the statistical properties of speech and noise to isolate one from another while multi-microphone techniques generally exploit spatial characteristics of the speaker and noise sources, often requiring precise microphone placement and calibration. This thesis proposes a novel and robust bi-microphone system aimed to fill the gap between the two extremes. Speech coherence between the two sound sources is exploited to supplement state of the art single microphone methods that use suppression rules and effective noise estimation techniques; the requirements on microphone directivity and placement are minimal, which enables rapid integration in current mobile devices. Subjective and objective results attest of the system's improved noise suppression compared to state of the art techniques in low signal to noise ratios.
ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor Roch Lefebvre, without whom nothing of this would have been possible. Despite his overbooked schedule, he still managed to find some time to give me cues, hints and directions at crucial times in my ongoing research. Along with Jean Rouat, I would like to thank both professors for giving me unique chances and encouragements in publishing my works, while supporting me financially in presenting my research’s findings; this is a chance given to few students and which deserves particular recognition.

I’m also extremely grateful to my parents for giving me the opportunities in life to go as far in my academic studies; I wouldn’t have achieved half of all this if it were not for their constant moral and physical support.

Next, thanks to all my friends for helping me in keeping a certain degree of sanity through the last few years. Be it those late coffee breaks with Jimmy or those afternoon baby foot furies with Frederick, I remain convinced that one cannot accomplish a master’s research without relaxing or switching it off at times.

Finally, and most importantly, thanks to Jamel. Thanks for your emotional support since the beginning of my undergraduate studies, and thanks for enduring my too frequent geeky talks and mumblings. Mostly, thanks for understanding me and still loving me despite everything and the huge sacrifices made in order to complete this work.

Again, thank you all.
# TABLE OF CONTENTS

1 INTRODUCTION .................................................................................................................. 1  
1.1 Applications ................................................................................................................. 2  
1.2 State of the art ............................................................................................................. 3  
1.2.1 Single microphone speech enhancement .............................................................. 3  
1.2.2 Multi-microphone speech enhancement ................................................................. 4  
1.2.3 Perceptual speech enhancement ............................................................................. 5  
1.3 Contribution .................................................................................................................. 5  
1.4 Thesis organization ..................................................................................................... 6  

2 SINGLE MICROPHONE SPEECH ENHANCEMENT .......................................................... 7  
2.1 Overview ...................................................................................................................... 7  
2.2 Analysis and synthesis framework ............................................................................. 8  
2.3 Choice of analysis/synthesis function ...................................................................... 8  
2.4 Wavelet modeling ..................................................................................................... 9  
2.5 Subspace modeling ................................................................................................... 11  
2.6 Autoregressive modeling .......................................................................................... 14  
2.6.1 Iterative Wiener filtering ..................................................................................... 15  
2.6.2 Kalman Filtering ................................................................................................ 17  
2.7 Spectral modeling ..................................................................................................... 17  
2.7.1 Suppression rules ................................................................................................. 19  
2.7.2 Noise estimation .................................................................................................. 24  
2.7.3 Artifacts and distortions ....................................................................................... 27  
2.8 Summary .................................................................................................................... 29  

3 MULTI-MICROPHONE SPEECH ENHANCEMENT .......................................................... 30  
3.1 Overview ...................................................................................................................... 30  
3.2 Fixed Beamforming ................................................................................................. 32  
3.2.1 Microphone characteristics ................................................................................. 33  
3.2.2 Delay and sum beamforming ............................................................................. 34  
3.2.3 Filter and sum beamforming ............................................................................. 36
3.3 Adaptive Beamforming ................................................................. 37
3.3.1 Generalized Sidelobe Canceller ................................................ 37
3.3.2 Multi-channel Wiener Filter ..................................................... 41
3.4 Coherence based filtering ............................................................ 43
3.4.1 Concepts and definitions ....................................................... 43
3.4.2 Connection with ANC performance ........................................ 45
3.4.3 Post-filtering ................................................................. 45
3.5 Other Methods ........................................................................... 49
3.5.1 Blind Source Separation ...................................................... 50
3.5.2 Phase-Error based Filtering .................................................. 50
3.6 Summary ...................................................................................... 52

4 PERCEPTUAL SPEECH ENHANCEMENT ......................................... 53
4.1 The PEAQ masking model ......................................................... 53
4.2 Application to speech enhancement .......................................... 55
4.2.1 Tsoukala & Soulodre’s method .............................................. 55
4.2.2 Virag’s method ................................................................. 56
4.2.3 Musical noise suppression .................................................... 57
4.3 Summary ...................................................................................... 57

5 PROPOSED SYSTEM ........................................................................ 58
5.1 Overview ...................................................................................... 58
5.2 Single microphone speech enhancement .................................... 58
5.3 Multi-microphone post-filter ..................................................... 60

6 EXPERIMENTS & RESULTS .......................................................... 63
6.1 Objective evaluation ................................................................. 63
6.2 Subjective evaluation ............................................................... 66
6.3 Performance and experimental setup ......................................... 68

7 CONCLUSION ................................................................................. 70
7.1 Summary ...................................................................................... 70
7.2 Further research directions ....................................................... 72
7.2.1 Alternative microphone placement ...................................... 72
LIST OF FIGURES

Figure 2.1 Popular analysis/synthesis window functions ........................................... 9
Figure 2.2 Time/Frequency plane of Fourier and Wavelet basis .............................. 9
Figure 2.3 Example of a wavelet packet tree .............................................................. 10
Figure 2.4 Speech segment and its associated Scree graph ........................................ 12
Figure 2.5 Iterative AR Speech Enhancement Method ............................................... 15
Figure 2.6 Speech enhancement by spectral modeling ............................................... 18
Figure 2.7 Attenuation curves for different suppression rules .................................... 23
Figure 2.8 Attenuation curves for the EMSR ............................................................... 24
Figure 2.9 Noise PSD Estimation with a VAD ............................................................ 25
Figure 2.10 MCRA Noise Estimation Technique ....................................................... 26
Figure 2.11 Sample ratio distributions for MCRA noise estimation ................................ 27
Figure 2.12 Sample musical noise ............................................................................. 28
Figure 3.1 General representation of beamforming ..................................................... 32
Figure 3.2 Microphone array far-field assumption ..................................................... 33
Figure 3.3 Common microphone directionality patterns .............................................. 34
Figure 3.4 Delay and sum beamformer ...................................................................... 35
Figure 3.5 Directivity of the delay and sum beamformer .......................................... 36
Figure 3.6 Structure of the Generalized Sidelobe Canceller ........................................ 38
Figure 3.7 Noise reduction performance of the GSC .................................................. 40
Figure 3.8 Structure of the Multichannel Wiener Filter .............................................. 41
Figure 3.9 SNR Improvements of the MWF ............................................................... 42
Figure 3.10 Theoretical and measured diffuse noise coherence .................................. 47
Figure 3.11 $H_{MSC}$ as a function of the MSC ............................................................ 48
Figure 3.12 $H_{MSC}$ in dB as a function of the input SNRs ......................................... 49
Figure 4.1 PEAQ masking model ............................................................................. 53
Figure 4.2 Sample PEAQ masking threshold ............................................................ 55
Figure 5.1 Proposed system overview ...................................................................... 58
Figure 5.2 Proposed single microphone speech enhancement ................................... 59
Figure 5.3 Proposed multi-microphone post-filter ..................................................... 61
### ACRONYMS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANC</td>
<td>Adaptive Noise Canceller</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>BSS</td>
<td>Blind Source Separation</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>EMSR</td>
<td>Ephraim &amp; Malah Suppression Rule</td>
</tr>
<tr>
<td>EVRC</td>
<td>Enhanced Variable Rate Codec</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>GSC</td>
<td>Generalized Sidelobe Canceller</td>
</tr>
<tr>
<td>GSS</td>
<td>Generalized Spectral Subtraction</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>KLT</td>
<td>Karhunen-Loeve Transform</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Prediction Coefficient</td>
</tr>
<tr>
<td>LSD</td>
<td>Log-Spectral Distortion</td>
</tr>
<tr>
<td>LP-LSD</td>
<td>Linear Prediction LSD</td>
</tr>
<tr>
<td>MSC</td>
<td>Mean Square Coherence</td>
</tr>
<tr>
<td>MCRA</td>
<td>Minima Controlled Recursive Average</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>MWF</td>
<td>Multichannel Wiener Filter</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PHAT</td>
<td>Phase Transform</td>
</tr>
<tr>
<td>PEAQ</td>
<td>Perceptual Evaluation of Audio Quality</td>
</tr>
<tr>
<td>PSD</td>
<td>Power Spectral Density</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>SRT</td>
<td>Speech Recognition Threshold</td>
</tr>
<tr>
<td>TDOA</td>
<td>Time Delay of Arrival</td>
</tr>
<tr>
<td>TEO</td>
<td>Teager Energy Operator</td>
</tr>
<tr>
<td>VAD</td>
<td>Voice Activity Detector</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

Noise, n.: *Any signal that does not convey useful information.*

But how do we separate useful from useless information in a mix of both? This is the question that speech enhancement techniques seek to answer by trying to remove background acoustic noise (useless information) from a signal of interest, in our case speech (useful information). With the advent of technologies such as mobile telephony, we are faced with the problem of device usage in increasing acoustically adverse conditions; this is where speech enhancement comes into play. Car engine noise, crowd babble, fan hum and even other speakers are examples of interferences that will degrade clean speech produced by a speaker of interest; reducing this amount of added noise thus becomes a requirement in order to enable a proper, pleasant and intelligible communication.

Various tools are often used in the field to achieve this goal; to name a few, speech production models, statistical models, spatial filtering and perceptual criteria are examples of techniques and notions employed and often combined in speech enhancement systems. They can also be classified in 2 categories; single microphone (one microphone) and multi-microphone (>2 microphones) input systems. Single microphone systems rely on speech modeling and noise estimation techniques to identify and handle noise components of the input signal. Multi-microphone systems will generally exploit the spatial properties of sound and are traditionally implemented using carefully designed microphone arrays. Conversely, correlation of speech or noise between microphones as a different interpretation of spatial information can be used to extract additional cues in the optic of speech enhancement.

The subject of this thesis is to propose a bi-microphone speech enhancement system that will fill the gap between single microphone and complex multi-microphone methods. The aim is to complement state of the art single microphone techniques with additional information from a secondary microphone, as to prevent speech distortion and under or over noise suppression often inherent in those systems. Using recent noise estimation techniques, single microphone noise suppression rules and multi-microphone coherence post-filtering, we present a system
that outperforms single microphone methods while being computationally reasonable. Objective and subjective evaluation of the proposed algorithms are used to quantify the performance of the compared systems.

1.1 Applications

Most voice communication systems include, at one stage or another, a speech enhancement processing block. This processing aims to increase the SNR of the sampled speech, as to enable communication in noisier environments, or at least improve intelligibility and quality of the perceived speech. The most popular examples of systems that require and use speech enhancement technologies are:

1. **Mobile phones**: As a consequence of the ‘use-anywhere’ concept of mobile telephony, these devices are most often used in acoustically adverse environments. Thus, speech codecs used in wireless applications will frequently implement noise suppression algorithms on the encoder side of the codec. The EVRC standard for example [9] [4.1.2], estimates the SNR of the sampled speech to effectively construct a noise suppression filter.

2. **Hearing aids**: A major complaint of people having cochlear hearing loss is the difficulty of understanding what is spoken when background noise is present [64]. Studies show that people with normal hearing have an SRT (speech-reception threshold) of 50% at an SNR of -5 dB; that is, they can understand half of what is said at that SNR. People with hearing disabilities require an SNR of +5 dB on average to obtain the same SRT [14]. Noise suppression algorithms attempt to reach that increase needed in SNR.

3. **Voice recognition systems**: ASR (Automatic Speech Recognition) systems can greatly benefit from speech enhancement processing. State of the art ASR systems are implemented as HMM (Hidden Markov Model) state models driven by cepstral information derived from the input waveform; distortions in the input caused by ambient noise lead to inaccurate cepstral information which severely affects the
1 Introduction

performance of the system. As studies examining that specific application have shown [26], noise reduction can triple word recognition rates in low SNRs (-5 dB) and nearly double in higher SNRs (+10 dB).

4. In-vehicle communication systems: Be it hands-free car telephony systems or jet-fighter aircraft communication systems, engine noise has proven to severely degrade the quality of perceived speech [58]. In fact, aircraft cockpit speech SNR has been shown to be in the range of -5 dB to 0 dB, which is clearly inappropriate for communication of information in mission critical scenarios. Again, speech enhancement methods help in reducing background noise, thereby enhancing the speech SNR and enabling proper communication.

Numerous other applications also benefit from more general audio denoising; audio restoration, hiss removal, noise-cancelling headphones, industrial noise control, etc. The line between speech and audio noise suppression is often very thin and many concepts and theories are shared between the two fields.

1.2 State of the art

To cover all existing noise suppression methods would require several hundred pages. Since this document aims to give the reader a broad yet concise understanding of the research field, we will focus on the key milestones that have influenced the research avenues taken over the years up to what is considered state of the art by today's standards.

1.2.1 Single microphone speech enhancement

Since the complexities of the theories developed naturally followed the hardware on which they could be implemented, it is obvious that the noise suppression methods of the mid 70s did not bare today's computational requirements. On that line, the first speech enhancement algorithms were mainly single microphone, using only one microphone as an input source. Notions from adaptive filtering and statistics were used to derive the classic Wiener filter [78], whose suppression rule was later generalized as 'spectral subtraction' [6]. Several refinements were made to that method [4] by adding oversubtraction factors and introducing a
noise floor to control residual noise. Of equal importance were the optimal estimators derived by Ephraim & Malah [15, 16] that minimized respectively the MMSE of the spectral and log-spectral distance between the estimated and true speech signal amplitude. Laplacian and gamma speech distributions were recently analyzed and optimal estimators [49] show a small increase in performance compared to the ones previously derived.

Different means of estimating noise were also derived and analyzed using minimum statistics [46, 48] which can also be used with traditional spectral subtraction methods. While the previously mentioned methods transform the signal in the frequency domain, other techniques use subspace [17, 29, 53] wavelet [3, 13] and auto-regressive models [26, 40]. Kalman filters [35] were also applied to speech enhancement [56] to track and predict the state vector of the AR model. Even neural networks [18, 76] occupy a small place in the field of speech enhancement.

1.2.2 Multi-microphone speech enhancement

Following the exponential increase in performance of integrated circuits during the mid 80s, multi-microphone processing became a viable solution for real-time processing of signals. Fixed beamforming [72] enabled the use of spatial information which is nonexistent in single microphone systems. The microphone array can steer its response in the direction of the desired speaker and attenuate noise coming from other directions. Extensions to the fixed beamformer combining adaptive filtering and spatial filtering led to the notion of adaptive beamforming, whose main implementations were in the form of the GSC (Generalized Sidelobe Canceller) and the MWF (Multi-microphone Wiener Filter) [44, 64]; adaptation leads to better immunity against microphone mismatch and enables suppression of moving interferers.

Other approaches using the MSC (Mean Square Coherence) [24, 73], a measure derived from the cross-correlation, can extract a spectral profile of either noise or speech, depending on the sampling conditions and microphone positioning. BSS (Blind Source Separation) methods such as ICA (Independent Component Analysis) and PCA (Principal Component Analysis) have also been applied to the problem [57, 75], but have so far provided very limited
1 Introduction

improvement over other techniques. Yet another technique called phase-error based filtering has recently been proposed [1, 25, 62] for dual-microphone speech enhancement systems. Signal subspace approaches have also been extended to benefit from multi-microphone information [28].

1.2.3 Perceptual speech enhancement

More recently, research has been steered towards understanding the mechanics of auditory perception. It has been observed that both the human ear and the brain play a crucial role in the transformation of sound waves into perceived sound. Research has mainly been focused on modelling the ear rather than the brain, as the inner workings of the latter are still far from being understood.

Consequently, it has been noted that both spectral and temporal masking occur physically within the inner ear. Inner ear spectral models [31, 63] have thus been designed to simulate the effects of masking, as a mean of determining what is heard and what is unheard, or masked, in an audio signal. This led to the development of perceptual audio codecs [55] such as MPEG1 Layer3 and AAC, where fewer or no bits are spent to code masked information. These notions have been transposed to the field of speech enhancement such as to minimize the distortion of speech in noise suppression algorithms. Specifically, it has been suggested to attenuate only the unmasked noise just under the masking threshold such that it cannot be heard [70], instead of applying perceptually blind suppression rules that are prone to speech distortion. Another suggested approach is to control the oversubtraction and noise flooring factors of the GSS (Generalized Spectral Subtraction) based on the calculated perceptual masking threshold [74]. Musical noise can also be detected and attenuated by finding unmasked tones in the filtered output that are absent in the noisy input signal [30].

1.3 Contribution

This thesis contributes to the field of speech enhancement in several ways, by:
1. suggesting a **novel bi-microphone noise suppression system** based on minimum statistics noise estimation and a weighted coherence-based post-filter, appropriate for bi-microphone hand-held devices

2. using state of the art **MCRA noise estimation techniques** with several single channel noise **suppression rules**, while presenting objective and subjective performance comparisons of the given algorithms

3. focusing on noise reduction of **real noisy speech samples** instead of simulated noise scenarios; weighting of the post-filter with a theoretical noise field coherence model that corresponds to noise observations

4. proposing implementations **independent as much of input parameters**; algorithms adapt to different sampling frequency and windowing settings, as care is taken to propose time constant proportional factors instead of fixed discrete values

### 1.4 Thesis organization

Single microphone techniques and the basic concepts of speech enhancement are presented in Chapter 2. Various multi-microphone noise suppression strategies are then explained in Chapter 3. Next, we explain in Chapter 4 a model based on auditory perception, and how it could be applied to our problem. A new bi-microphone speech enhancement system is then proposed in Chapter 5, as well as its performance compared to other state of the art techniques in Chapter 6. A brief wrap-up and possible other research avenues are finally addressed in the conclusion, Chapter 7.

It is also of note that the reader can refer itself to the table of acronyms as a reference for commonly used expressions.
CHAPTER 2
SINGLE MICROPHONE SPEECH ENHANCEMENT

2.1 Overview

Single microphone speech enhancement techniques essentially try to extract the speech part of a ‘speech + noise’ mix. The basic model on which most single microphone techniques are based is described by:

\[ x[n] = s[n] + n[n] \] (2.1)

where \( s[n] \) is the clean speech signal, \( n[n] \) is the additive noise we are trying to remove and \( x[n] \) is the contaminated speech signal that will be used for processing. Since no information is known about the particular noise \( n[n] \), there are two general approaches that are taken by speech enhancement algorithms:

1. model the underlying speech signal \( s[n] \), such that an estimate \( \hat{s}[n] \) of the clean speech can be made from \( x[n] \)

2. model the underlying noise signal \( n[n] \), such that its estimate \( \hat{n}[n] \) can be somehow removed from the noisy mix \( x[n] \).

This modeling and estimation of either speech or noise components is rarely done in the time-domain since it does not convey useful information about neither their periodic characteristics (in speech) nor statistical stationarity (in noise). Instead, short contiguous frames of \( x[n] \) called windows are extracted and projected in another domain, such as the wavelet, subspace, auto-regressive or spectral domain, where speech enhancement rules are applied. The enhanced speech windows are later transformed back to the time domain where they can be concatenated to produce the final clean speech estimate. This operation and the most popular models used are described in details in the following discussion.
2.2 Analysis and synthesis framework

To exploit the fact that speech can be considered stationary under short time periods [41], \( x[n] \) is cut into short windows of length 5-50 ms, where 32 ms is generally used as it produces lengths in samples that are powers of 2 (512 samples at \( f_s = 16 \) kHz) thus speeding up the calculation of numerous domain transforms. Concrete evidence justify the choice of such a window length, as the average length of a speech phoneme is known to be around 30 ms for stop-consonants to 80 ms for vowels [61]. This windowing operation is described by (2.2):

\[
x_k[n] = x[kS + n]w_k[n], \text{ for } n = 0 \ldots N - 1
\]

where \( k \) is the window index, \( S \) is the window step size, \( N \) is the window length in samples and \( w_k[n] \) is a length \( N \) analysis window function. To avoid artefacts caused by window boundaries, a certain amount of overlap, most often 50%, is introduced during windowing. The computation of \( S \) and \( N \) for an arbitrary \( f_s \) and overlap ratio, is explained in Appendix 8.1. This step is referred as the analysis part of processing.

Processing is then performed on those windows of \( x_k[n] \), such as to compute an estimate \( \hat{s}_k[n] \) of \( s_k[n] \). Overlap-add is used to reconstruct the final output estimate \( \hat{s}[n] \) as per eq. (2.3).

\[
\hat{s}[n] = \sum_{k,m} \hat{s}_k[m] \cdot w_k[m], \text{ for } n = kS + m \text{ and } m = 0 \ldots N - 1
\]

2.3 Choice of analysis/synthesis function

One requirement of the analysis and synthesis window combination used is that they add up to unity when multiplied and overlapped in time. This is why, to simplify matters, both windows are often chosen as the square root of a main, common window \( w[n] \):

\[
w_a[n] = w_s[n] = \sqrt{w[n]}
\]
Common window functions include the square, Hamming, Hanning and Bartlett windows, as defined in Appendix 8.2. The square root of those windows are illustrated in Figure 2.1 as well as their corresponding frequency response, for \( f_s = 16 \) kHz and 32 ms long windows.

![Figure 2.1 Popular analysis/synthesis window functions](image)

The actual shape of the window has been shown to have little effect on the performance of speech enhancement algorithms [41]. Since the root Hanning window has both great frequency resolution and high sidelobe attenuation, it will be used for both analysis and synthesis of our short-time speech enhancement algorithms.

### 2.4 Wavelet modeling

One proposed attempt at modeling speech is through the use of wavelet bases [3, 13]. The main advantage of the wavelet transform over the much popular Fourier transform is in the non-periodic nature of its basis functions; they can differ in length and are thus more apt at modeling speech transients and short, isolated events (Figure 2.2).

![Figure 2.2 Time/Frequency plane of Fourier and Wavelet basis](image)
The wavelet decomposition essentially calculates a projection of scaled and translated versions of a wavelet mother function on $x_n$, the noisy speech signal. Popular wavelet mother functions used for audio and image denoising [45] include the Daubechies, Symlet and Coiflet wavelet families. The wavelet decomposition can also be viewed as a QMF filter bank, where the orthogonal low-pass and high-pass analysis filters are derived from the mother wavelet function. This decomposition produces what is called a wavelet packet tree (Figure 2.3), where each node represents the output of one QMF filter at a given decimation level $l$. There is redundancy in the representation, such that each node at level $l+1$ contains half the coefficient than its parent at level $l$ while the total number of coefficients at each level remains constant; that is, any non-redundant combination of nodes can be used to perfectly reconstruct the original waveform. One such combination, called a wavelet tree, is illustrated (greyed out) in Figure 2.3.

![Wavelet Packet Tree Diagram](image)

**Figure 2.3 Example of a wavelet packet tree**

Denoising is done by thresholding the selected wavelet coefficients, on the assumption that the wavelet basis better represents the speech components than the noisy ones, such that those noisy coefficients shall be the smallest and thus easily identified. This is referred in the literature as wavelet shrinkage, and the challenge lies in determining:

1. the optimal wavelet tree
2. the thresholds used at each node.

Coifman [11] tackled the issue of a ‘best-tree’ selection algorithm and proposed a popular method that consists in minimizing the Shannon entropy of the selected coefficients. On the other hand, the universal threshold $\lambda$ proposed by Donoho [13] is often used in speech
enhancement applications. More recently, it has also been suggested by Bahoura [3] to use a measure proportional to the TEO (Teager Energy Operator) of the noisy signal as a time-adapted threshold.

However, recent experiments have only shown small increases in SNR (2 to 10 dB) in noisy input conditions (-10 to 5 dB). The perceptual improvements obtained are not significant compared to computationally simpler spectral methods that will be presented later.

2.5 Subspace modeling

The idea behind subspace decomposition consists in extracting an orthogonal vector basis of a signal that has maximum energy compaction. Conversely, this decomposition is said to be its optimal low-rank representation. The problem can be formulated as:

\[
\min_{\theta, \mathbf{v}} \mathbb{E}\left\{ \left\| \mathbf{x} - \sum_{i=1}^{K} \theta_i \mathbf{v}_i \right\|^2 \right\} \tag{2.5}
\]

where \( \mathbf{x} = x_i \mathbf{[n]} \) (see (2.1)). The problem is to find the scalars \( \theta_i \) and the vector basis \( \mathbf{v}_i \) for \( i = 1 \ldots K \), such that \( K < N \) where \( \mathbf{x} \) is \( N \)-dimensional. This is known as the Karhunen-Loève Transform (KLT) and its solution is found by solving the eigenvectors and eigenvalues of the autocorrelation matrix of \( \mathbf{x} \), stated as:

\[
\mathbf{v}_i \mathbf{R}_{xx} = \theta_i \mathbf{v}_i \tag{2.6}
\]

The idea of reducing the dimensionality of \( \mathbf{x} \) is supported by observing the ‘eigenvalue spectra’ of a speech signal over time, also known as its Scree graph [32] (Figure 2.4).
2.5 Subspace Modeling

Figure 2.4 Speech segment and its associated Scree graph

This Scree graph was computed for short windows of 512 samples; as can be seen, the first 20 eigenvectors contain most of the energy present in the speech segments. It has been observed that fewer vectors are required for voiced segments (vowels, highly periodic) while a higher order model is required for unvoiced segments (consonants, less correlated waveforms).

A formal framework was devised in [16] to analyze the noise reduction problem under the signal subspace model. First, the speech segment \( s = s_\nu[n] \) was assumed to originate from a linear production model of order \( K \):

\[
s = \sum_{i=1}^{K} a_i v_i
\]  

(2.7)

The exact linear model (Fourier, autoregressive, etc.) is of no importance; however, the fact that \( K < N \) is (see Figure 2.4), as it implies that the covariance matrix of \( s \) will have null eigenvalues. The noise \( n = n_\nu[n] \) is assumed to be white, even though a pre-whitening filter based on the estimation of \( R_{nn}[n] \) is presented to handle the coloured noise cases. As such \( R_{xx}[n] \) can be expanded as (omitting sample indexes):

\[
R_{xx} = R_{xx} + R_{nn} = R_{xx} + \sigma_n^2 I
\]  

(2.8)

Writing the eigenvalue decomposition of (2.8) as \( R_{xx} = U\Lambda_s U^T \), the eigenvalues in the diagonal of \( \Lambda_s \) can be expanded as (see [16]):
\begin{align*}
\lambda_{i,j} &= \begin{cases} 
\lambda_{i,j} + \sigma_n^2 & \text{for } i = 1 \ldots K \\
\sigma_n^2 & \text{for } i = K + 1 \ldots N 
\end{cases} 
\end{align*} \tag{2.9}

The basic idea is then to estimate the order \( K \) of \( s \), such that the corresponding eigenvectors of \( U \), called the 'principal eigenvectors' can be extracted. Several order estimators minimizing different criteria such as distortion, entropy, and over/under estimation can be derived to deduce the order of \( s \) from a noisy eigenvector decomposition of \( R_{ss} \). For example, the theoretical estimator \cite{36} \( U_1 \) forms a \( N \times K \) matrix defined by:

\[ U_1 = \{ u_i : \lambda_{i,j} > \sigma_n^2 \} \tag{2.10} \]

This estimator ensures that poorly estimated eigenvalues and eigenvectors are not used to represent the signal. Even though it is computationally efficient, it is not optimal for full-rank signals and does not attempt to trade-off signal distortion and noise removal. The principal eigenvectors \( U_1 \) can then be used to construct a projector \( H \) of the noisy speech \( \hat{s} \) into the signal subspace. The simplest projection estimator that minimizes the Euclidian distance between \( x \) and \( s \) is the linear least square estimator, as defined by:

\[ \hat{s} = Hx = U_1U_1^Hx \tag{2.11} \]

Other projection estimators (Minimum Variance, Time Domain Constrained, etc.) can be used to minimize other criteria; perceptual elements can also be incorporated \cite{36} to focus on minimizing perceived noise only. Finally, overlap-add as defined in Section 2.2 is used to synthesize \( \hat{s} \) into \( \hat{s}[n] \).

The main drawback of signal subspace methods is their high computational complexity; eigenvector decomposition is very expensive. Also, the pre-whitening of the noisy signal requires knowledge or estimation of the noise auto-correlation matrix which is a non-trivial operation. In fact, recent experiments in signal subspace speech enhancement including auditory masking effects \cite{36} have only shown results for synthetically mixed samples where \( R_{ss} \) was known in advance. Even then, the system performed worst in coloured noise than in the white noise scenarios. For these reasons, other, simpler methods are generally preferred in the field of speech enhancement.
2.6 Autoregressive modeling

Autoregressive modeling consists in modeling speech production as the response of a linear system, the vocal tract, to either a periodic or noise-like excitation during voiced and unvoiced segments respectively. A commonly used and physically sound model for the vocal tract is the all-pole model $V(z)$ in eq. (2.12), as proposed by [40].

$$S(z) = U(z) \cdot V(z) = \frac{U(z)}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{U(z)}{\prod_{k=1}^{p} (1 - p_k z^{-k})}$$  \hspace{1cm} (2.12)

The signals are expressed in the $Z$-transform domain, where $S(z) = Z \{ s[n] \}$ and $X(z) = S(z) + N(z)$. Here, $U(z)$ is normally referred as the process' innovation signal, and is usually modeled by isolated impulses during voiced segments and random noise during unvoiced segments. Nonetheless, the exact form of $u[n]$ has little impact [22] and as such can be modeled as random white noise without having any significant influence on the estimation procedure of the all-pole model. From eq. (2.12), $s[n]$ can thus be represented in the time domain by:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + u[n]$$  \hspace{1cm} (2.13)

The challenge lies in estimating the $a_k$ parameters from the noisy speech observations $x[n]$. For that matter, two approaches are generally used:

1. Iterative Wiener filtering with auto-regressive modeling
2. State-space model with Kalman filtering

We will discuss and explain the two techniques, as well as the basic theory behind the Wiener filter which is going to come handy in the next sections on spectral modeling methods.
2.6.1 Iterative Wiener filtering

The iterative Wiener AR speech enhancement procedure initially proposed by [40, 41] is illustrated in Figure 2.5.

![Diagram](image)

*Figure 2.5 Iterative AR Speech Enhancement Method*

The general idea is to model the speech PSD $P_s(f)$ using the auto-regressive model described by (2.13), and to iteratively apply a Wiener filter until a fixed number $n$ of iterations or until a distortion criteria falls under a certain threshold [41]. The noise PSD $P_n(f)$ has to be approximated with noise estimation techniques (see Section 2.7.2), an issue which can affect the performance of the algorithm.

2.6.1.1 Wiener Filter

The general Wiener filter $h[k]$ seeks to find a linear MMSE (Minimum Mean Square Error) estimate of a desired signal $d[n]$ given an observation $x[n]$. It is derived in Appendix 8.3 and restated here for convenience, substituting $d[n] \rightarrow s[n]$, as our desired signal is the clean speech signal:

$$\sum_{k=0}^{N-1} h[k] R_{x}[i-k] = R_{x}[i], \quad i = 0...N-1$$  \hspace{1cm} (2.14)
Since \( s[n] \) and \( n[n] \) are assumed to be uncorrelated,

\[
R_{ss}[n] = R_s[n] \\
R_{sn}[n] = R_s[n] + R_m[n]
\]  

(2.15)

Knowing that the Fourier transform of a cross-correlation \( R_{xy}[n] \) is the cross PSD (Power Spectral Density) \( P_{xy}(f) \), we can rewrite (2.14) as:

\[
H(f) = \frac{P_{x}[f]}{P_{x}(f) + P_{m}(f)}
\]  

(2.16)

which is the form used in Figure 2.5.

2.6.1.2 Relation between Wiener filtering and the LPC coefficients

The last point to resolve is the estimation of the \( a_k \) coefficients in (2.13) used to model \( P_{x}[w] \). This is done again using the general Wiener filter definition, substituting \( d[n] \rightarrow s[n] \), \( x[n-k] \rightarrow s[n-k] \) and \( h[k] \rightarrow a_k \) for \( k = 1 \ldots N \).

\[
\sum_{k=1}^{N} a_k R_{ss}[i-k] = R_{s}[i], \quad i = 0 \ldots N-1
\]  

(2.17)

We arrive at eq. (2.17), known as the Yule-Walker equation, which can be optimally solved using the Levinson-Durbin recursion that accounts for the redundancy of \( R_{ss}[n] \) on both sides of the equation.

The iterative technique presented above performs well, but it has been observed that the poles of the AR spectrum move rapidly from one frame to another. Hansen [26] suggested to impose inter-frame spectral continuity constraints, such that the PSD \( P_{x}(f) \) computed at iteration \( k \) isn’t too different from the one at iteration \( k+1 \). This is motivated by the fact that the characteristics of the vocal tract do not change drastically from one frame to another; small reduction in speech distortion has been observed using this technique.
2.6.2 Kalman Filtering

Another mean of solving the $a_k$ coefficients in (2.13) is through the use of Kalman filtering [37, 56]. It has been proven that solutions converge to the MMSE state of the optimum Wiener filter as $k$, the number of windows processed, tends to infinity [68]. Without delving into the details of Kalman filtering (refer to [35] for implementation details), the algorithm works by alternating between predict and correct steps of the state vector of the coefficients.

$$s_k = a^T s_{k-1} + K_k \left( x[n] - a^T s_{k-1} \right) \quad (2.18)$$

The state vector that includes the time domain estimate $\hat{s}[n-k]$ are first predicted a priori using the linear AR model in $a^T$ (written in bold character) and this prediction is then corrected using the observation $x[n]$ and the calculated Kalman gain $K_k$. This gain requires calculation of the auto-correlation matrix $R_\alpha[n]$ [56].

While auto-regressive methods have shown appreciable 5 to 10 dB improvements in SNR, their main appreciable gains have been in the recognition accuracy of ASR systems. No clear intelligibility improvements were noticed with this modeling method [37], while it was also noticed that it often required parameter tuning; simpler spectral methods were found to introduce less speech distortion in subjective evaluation.

2.7 Spectral modeling

The most successful speech enhancement methods yet applied to our problem are based on spectral modeling. One might wonder why it has attracted much of research interest in the last 30 years, instead of say wavelet or signal subspace methods. Several answers stem from the properties and characteristics of the spectral bases:

1. The bases are signal independent (those of signal subspace are not): this makes formal analysis of suppression rules and residual error a lot easier.
2. The bases are periodic (wavelets are not): this simplifies a lot deterministic and stochastic modeling of the Fourier coefficients; moreover, it offers a good representation of formants and speech characteristics.

3. Fast transforms have existed for long (by use of the FFT-based methods): real-time Fourier Transform were conceivable in the mid-80’s, while KLT and Wavelet transformation were not.

4. Auditory models, initially for use in audio compression applications, have been developed in the spectral domain; they are now used in speech enhancement methods.

The methods work as follows: they first transpose the time domain signal into the frequency domain by use of the Discrete Fourier Transform, a suppression rule is then applied, and the signal is finally converted back to its time domain representation by the inverse Discrete Fourier Transform. The general procedure is illustrated in Figure 2.6.

![Figure 2.6 Speech enhancement by spectral modeling](image)

We recall the well-known DFT:

\[ \mathcal{F}(x[n]) = X[m] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi mn/N}, \text{ for } m = 0 \ldots N-1 \]  
\[(2.19)\]

and its inverse, the inverse DFT:

\[ \mathcal{F}^{-1}(X[m]) = x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[m] \cdot e^{j2\pi km/N}, \text{ for } n = 0 \ldots N-1 \]  
\[(2.20)\]
Optimized versions of the DFT known generally as FFT-based methods exist and are used in practically all implementations. Thereafter, a suppression rule in the form of a frequency-domain filter \( H[m] \) is applied:

\[
\hat{S}[m] = H[m]X[m]
\]  

(2.21)

This suppression rule aims to attenuate noisy frequency bands, while minimizing speech distortion; this is the common trade-off that speech enhancement algorithms have to deal with.

We will now present different suppression rules \( H(f) \), how they were derived and how they compare against each other; the discussion will be carried on in the continuous frequency domain \( f \), while practical implementations obviously use the discrete frequency domain \( m \) discussed above. We will then explain different noise estimation techniques required for the application of those suppression rules, as well as the popular 'musical noise' problem which is most characteristic of spectral modeling methods.

2.7.1 Suppression rules

2.7.1.1 Wiener Filter

The first suppression rule is the Wiener filter \([78]\) and was already derived in Section 2.6.1.1. This gain function was obtained while minimizing the MSE between \( \hat{S}(f) \) and \( S(f) \), taking into account both phase and amplitude. Rewriting (2.16) with discrete frequency bands and dividing the right-hand side by \( P_{\text{na}}(f) \) we obtain:

\[
H_w(f) = \frac{\text{SNR}_{\text{post}}(f)}{\text{SNR}_{\text{post}}(f) + 1} 
\]  

(2.22)

where \( \text{SNR}_{\text{post}} \) is an approximation of the real \( \text{SNR} = P_{s} / P_{n} \) and defined as the a posteriori SNR. It is termed a posteriori, as opposed to the a priori SNR that we will see later, because it is evaluated after a new sample of \( X(f) \) has been observed. It is an instantaneous
estimation of the SNR that only considers the current spectral values. It is obtained by applying the approximation $E\{P_n(f)\} \approx |X(f)|^2$ and can be estimated as follows:

$$
SNR_{post}(f) = \frac{|X(f)|^2}{P_n(f)} - 1
$$

(2.23)

Putting (2.23) in (2.22), $H_p(f)$ can be rewritten in terms of $|X(f)|^2$ and $P_n(f)$, which highlights its 'noise subtraction' characteristic:

$$
H_p(f) = \frac{|X(f)|^2 - P_n(f)}{|X(f)|^2} = 1 - \frac{P_n(f)}{|X(f)|^2}
$$

(2.24)

2.7.1.2 Power subtraction

Since it is well known that the perception of speech is almost phase insensitive [79], one idea was to estimate $|S(f)|$ rather than $S(f)$, and append the noisy phase to the estimated $\hat{S}(f)$. A **maximum likelihood** estimator was derived for $|S(f)|$ (different from a MSE estimator) in [50], assuming Gaussian distribution models of $S(f)$ and $N(f)$, and obtained as (under its various forms):

$$
H_p(f) = \sqrt{\frac{SNR_{post}(f)}{SNR_{post}(f) + 1}}
$$

$$
= \sqrt{1 - \frac{P_n(f)}{|X(f)|^2}}
$$

(2.25)

The power subtraction rule ends up being the square root of the Wiener filter.

2.7.1.3 Generalized spectral subtraction

Berouti [4] suggested to generalize the subtraction rule to include a variable exponent, as:
2. SINGLE MICROPHONE SPEECH ENHANCEMENT

\[ H_{SS}(f) = \left( 1 - \frac{M_{ss}(f)^2}{|X(f)|^2} \right)^{\frac{1}{2}} = \left( 1 - \frac{1}{(\text{SNR}_{post}(f) + 1)^{\gamma/2}} \right)^{\frac{1}{\gamma}} \]  (2.26)

where \( M_{ss}(f) = E[|N(f)|^2] \). As such, \( H_{p}(f) = \left[ H_{SS}(f) \right]_{\gamma = 2} \) is the power spectral subtraction rule and \( \left[ H_{SS}(f) \right]_{\gamma = 1} \) is called magnitude spectral subtraction.

2.7.1.4 Maximum Likelihood Envelope Estimator

Another **maximum likelihood** estimator of \( |S(f)| \) was derived in [50], but under the assumption this time that \( S(f) \) is a **deterministic** complex exponential of fixed amplitude and phase (hence the term ‘envelope estimator’), and that \( N(f) \) still follows a Gaussian distribution:

\[ H_{MLEE}(f) = \frac{1}{2} + \frac{1}{2} \sqrt{\frac{\text{SNR}_{post}(f)}{\text{SNR}_{post}(f) + 1}} \]

\[ = \frac{1}{2} + \frac{1}{2} \sqrt{1 - \frac{P_{ss}(f)}{|X(f)|^2}} \]  (2.27)

2.7.1.5 Ephraim & Malah Suppression Rules

A formal effort was made by Ephraim & Malah [15] to derive a **MMSE** estimator of \( |\hat{S}(f)| \) (similar to the Wiener filter, but only in amplitude) under Gaussian models of \( S(f) \) and \( N(f) \). The following results were obtained, omitting frequency indexes for simplicity (as formulated by [7]):

\[ H_{EMSE}(f) = \frac{\sqrt{\pi}}{2} \sqrt{\frac{1}{1 + \text{SNR}_{post}} \left( \frac{\text{SNR}_{prio}}{1 + \text{SNR}_{prio}} \right)} \times M \left( 1 + \text{SNR}_{post} \left( \frac{\text{SNR}_{prio}}{1 + \text{SNR}_{prio}} \right) \right) \]  (2.28)

where:

\[ M(x) = \left[ \exp\left( -\frac{x}{2} \right) \right] \left[ (1 + x)I_0\left( \frac{x}{2} \right) + x \cdot I_1\left( \frac{x}{2} \right) \right] \]  (2.29)
and $I_n(\cdot)$ is the $n^{th}$ order modified Bessel function, implemented by \texttt{besseli(nu,x)} in Matlab. $\text{SNR}_{\text{post}}(f)$ is as always defined by (2.23), while $\text{SNR}_{\text{prio}}(f)$, the \textit{a priori} SNR, is defined by:

$$
\text{SNR}_{\text{prio}}(f) = \frac{P_n(f)}{P_m(f)}. 
$$

(2.30)

$\text{SNR}_{\text{prio}}(f)$, as opposed to $\text{SNR}_{\text{post}}(f)$, is an estimation of the SNR that contains information from all the past frames. Since $P_n(f)$ is obviously unknown, $\text{SNR}_{\text{prio}}(f)$ must be estimated. The most widely used estimation was suggested by [15] in their original paper, and is termed the ‘decision directed approach’:

$$
\text{SNR}_{\text{prio,k}}(f) \approx \alpha \frac{|\tilde{S}_{k-1}(f)|^2}{P_{m,k-1}(f)} + (1-\alpha) \text{SNR}_{\text{post,k}}(f)
$$

(2.31)

It is, actually, a time-smoothed version of the SNR, where $\alpha$ is set very close to unity ($\alpha = 0.99$).

Another suppression rule was derived by the same authors [16], under the same principles, minimizing this time the MMSE of $\log|\tilde{S}(f)| - \log|S(f)|$. Omitting frequency indexes:

$$
H_{\log \text{EMSR}} = \frac{\text{SNR}_{\text{prio}}}{1 + \text{SNR}_{\text{prio}}} \exp \left( \frac{1}{2} \exp \int \left( 1 + \text{SNR}_{\text{post}} \left( \frac{\text{SNR}_{\text{prio}}}{1 + \text{SNR}_{\text{prio}}} \right) \right) \right)
$$

(2.32)

where

$$
\exp \int(x) = \int_{-\infty}^{\infty} \frac{e^{-x}}{x} \, dx
$$

(2.33)

can be implemented by the use of \texttt{erf(x)} in Matlab. This estimator has been shown to produce slightly better results than $H_{\text{EMSR}}(f)$, as auditory perception is sensitive in a logarithmic scale.
2.7.1.6 Comparison of suppression rules

The principal ground of comparison of suppression rules is their attenuation curve as a function of the input SNR. This is plotted in Figure 2.7 for rules depending on $SNR_{\text{post}}$ only.

![Attenuation curves for different suppression rules](image)

Figure 2.7 Attenuation curves for different suppression rules

We see that $H_{\text{MLEE}}$ provides very little attenuation for low input SNR; it is of little use if significant noise attenuation is sought. $H_{W}$ provides strong attenuation in low input SNR, and relatively low attenuation for higher (>5 dB) SNR. On the other hand, $H_{ss}$ and $H_{F} = [H_{ss} | \lambda = 2]$ offer a convenient way to customize the gain curve by varying $\lambda$.

The scenario is a little different for the EMSR, as it depends both on the a priori and a posteriori SNR. A plot of $H_{\text{EMSR}}$ for $SNR_{\text{pro}}$ versus the attenuation is presented, for different values of $SNR_{\text{post}}$. Since the curves for $H_{\log \text{EMSR}}$ are only within a few dB’s (<3 dB) of those of $H_{\text{EMSR}}$, we will only focus on analysis of the latter.
As highlighted by Cappé [7], the a priori SNR is the dominant parameter for the EMSR. The a posteriori SNR acts as a correction parameter whose influence is limited to the case where the a priori SNR is low. The behaviour of $\text{SNR}_{\text{post}}$ is counter intuitive at first; the larger $\text{SNR}_{\text{post}}$ is, the higher the attenuation. This is a consequence of the disagreement between $\text{SNR}_{\text{prio}}$ and $\text{SNR}_{\text{post}}$. This is a very useful property of this estimator and explains why it is less susceptible to musical noise than the other estimators.

2.7.2 Noise estimation

The previously discussed suppression rules all require the estimation of $\text{SNR}_{\text{post}}$. By eq. (2.23), this is done through the observed noisy speech power $|X(f)|^2$, and an estimate of the noise PSD $P_n(f)$ that still needs to be resolved. This noise estimation problem is unfortunately often overlooked by many research papers which only consider denoising of speech corrupted by ergodic generated white Gaussian noise, manually added to an original clean speech signal, and whose variance is known in advance.

The fact is that real life noise is dynamic both in amplitude and in frequency. Furthermore, it has been shown [71] that the estimated noise has a profound influence on the amount of
residual noise left by spectral subtraction techniques. Thus, noise estimation methods have been devised, and we illustrate in Figure 2.9 the basic principle under which they generally operate.

As shown above, the idea is to estimate, in each spectral bin, the noise power in noise-only periods. The design of a Voice Activity Detector (VAD) that will switch on and off noise estimation is non-trivial. The PSD estimation block is normally a simple first-order IIR filter that will smooth the power estimate over time.

Since there are so many noise estimation techniques and so many VADs based on totally different approaches (Markov Chains, Neural Networks, AR modeling, etc.), we will focus on a popular method proposed by Cohen [10] called “Noise Estimation by Minima Controlled Recursive Averaging (MCRA)”, extending on the works of Martin [47]. The technique uses minimum statistics of the noise spectrum and has been successfully applied to spectral enhancement of speech. It is also easy to implement, well documented, and can obviously be coupled with all the suppression rules previously introduced.

2.7.2.1 Minimum Statistics Noise Estimation

The idea behind recently introduced minimum statistics noise estimation techniques stems from the observation that in a particular frequency bin, noise can be estimated even during speech periods if the minima (i.e. the dips in amplitude) are tracked correctly. The MCRA algorithm proposed by Cohen is illustrated in Figure 2.10.
Figure 2.10 MCRA Noise Estimation Technique

The IIR blocks in Figure 2.10 are simply first-order IIR filters adjusted with a given time constant, as defined in Appendix 8.4; this avoids the use of fixed $\alpha$ parameters and thus adapts to sampling frequency, window overlap and window length.

In short, the technique consists in calculating a ratio $R$ between the instantaneous local energy and the minimum energy observed over a fixed number of past windows. The energy of the noisy spectrum is lightly smoothed both in time and in frequency before calculating the ratio which is used to determine speech presence. In fact, the detection of speech is based on the Bayes minimum-cost decision:

$$ R = \frac{p(R|H_1)}{p(R|H_0)} > \delta \rightarrow H_1 $$
$$ R = \frac{p(R|H_1)}{p(R|H_0)} < \delta \rightarrow H_0 $$

where $H_0$ and $H_1$ are the hypothesis of speech absence and presence respectively. This decision criterion is motivated by the observation of the distributions $p(R|H_0)$ and $p(R|H_1)$ for known speech samples. This is illustrated in Figure 2.11, taken from [10], for speech mixed with F16 cockpit noise.
The binary speech detector $I = R > \delta$ is then smoothed into a speech presence probability $p$, which is in turn smoothed into $a_{\text{noise}}$ to control the time constant of the PSD estimation. For example, when speech is present $I \to 1$ and $a_{\text{noise}} \to 1$ such that the noise estimate $P_{nn}[k]$ is not updated with observed (speech only) values of $|X[k]|^2$, and vice-versa.

2.7.3 Artifacts and distortions

2.7.3.1 Musical noise

The most annoying distortion caused by short time spectral enhancement techniques is the apparition of a tone-like residual noise, coined musical noise, in the speech estimate spectrum $\hat{S}(f)$. Even though the SNR is undoubtedly increased, the window-to-window approach causes subtraction of an averaged estimate noise power from an instantaneous, highly varying noisy speech spectrum. This musical noise is perceived as random isolated tones appearing and disappearing quickly; it is often more annoying than the original noise which we are trying to remove. Sample musical noise caused by the GSS suppression rule with $\gamma = 2$ (see Section 2.7.1.3) is illustrated in Figure 2.12.
Figure 2.12 Sample musical noise

Two solutions were proposed initially by Berouti [4] to overcome this effect of the GSS suppression rule. The first was to over subtract the noise estimate with a factor $\alpha$, such as to reduce the amplitude of those resulting peaks. The second was to introduce an adaptive flat noise floor (proportional to the estimated noise by a factor of $\beta$) that masks the resulting musical noise. As derived in the original paper, the resulting output spectrum for the case of magnitude spectral subtraction ($\gamma = 1$) can be expressed as:

$$|\hat{S}(f)| = \max \left( |X(f)| - \alpha \sqrt{P_{nn}(f)}, \beta \sqrt{P_{nn}(f)} \right)$$  \hspace{1cm} (2.35)

We recall that certain suppression rules suffer a lot less from the musical noise phenomenon. This is especially true for the EMSR and log-EMSR rules as they take into account discrepancies between instantaneous and mean changes in the estimated SNR.

2.7.3.2 Phase distortion

Another artefact is the phase distortion caused by taking the phase of the noisy speech for our speech estimate, that is $\angle \hat{S}(f) = \angle X(f)$. Assuming we were able to perfectly estimate the SNR and thus perfectly retrieve $|S(f)|$, we can make an experiment whereas we compute the speech estimate $\hat{S}(f) = |S(f)| e^{j\alpha X(f)}$. What is found is that under 0 dB of SNR, the resulting speech estimate has a ‘hoarse’ sounding voice; listen to the sample files in the accompanying
media package of this research [19]. As such, the assumption that the ear is phase insensitive does not hold for very noisy regions of the spectrum.

2.8 Summary

In this chapter, we have presented different single microphone speech enhancement methods. The short-time analysis/synthesis framework was presented, as well as the impact of its windowing operation on speech processing. Speech enhancement by wavelet, signal subspace and autoregressive methods was explained. More emphasis was put on simpler and proven spectral enhancement methods that often result in less speech distortion and more noise suppression due to better understanding of the spectral domain. We discussed and compared several suppression rules used in spectral enhancement techniques, as well as one recent and robust noise estimation algorithm; the Ephraim & Malah suppression rule were shown to attenuate the problem of musical noise. Finally, common artefacts and distortions caused by spectral enhancement were explained, as well as some solutions proposed for them.
CHAPTER 3
MULTI-MICROPHONE SPEECH ENHANCEMENT

3.1 Overview

As DSPs performance increased, mobile devices requiring noise reduction functions have forayed into the field of multichannel processing. Miniaturization of microphones [54] have paved the way for devices such as three microphone hearing aids [42] which aim to improve on the performance of single channel noise suppression techniques. Different noisy speech versions are captured at each microphone, in the aim of acquiring more information about the individual clean speech and noise components.

The signal model generally used can be formulated as follows:

\[ x_i[n] = h_i[k] * s[n] + n_i[n] \]
\[ x_2[n] = h_2[k] * s[n] + n_2[n] \]
\[ \cdots \]
\[ x_M[n] = h_M[k] * s[n] + n_M[n] \] (3.1)

where there are up to \( M \) microphones (i.e. \( i = 1 \ldots M \)), \( h_i[k] \) is the transfer functions from the source to microphone \( i \) and \( n_i[n] \) are the different noise sources. The model in (3.1) is also often rewritten in a more convenient matrix form as:

\[ X = H \times S + N \] (3.2)

where:

\[ X = [x_1[n] \ldots x_M[n]]^T \]
\[ H = [h_1[k] \ldots h_M[k]] \]
\[ S = [s_1[n] \ldots s_M[n]]^T \]
\[ N = [n_1[n] \ldots n_M[n]]^T \] (3.3)

Even though we will adopt the \( M \) microphone multi-channel notation, we will focus on the bi-microphone case as this is the application of this research project, and also because the concepts are simpler to explain and demonstrate in that particular framework.
That being said, multi-microphone techniques can thus exploit spatial information in addition to spectral and temporal information as opposed to single microphone methods presented in Chapter 2. Through careful combination of the observed signals $x_i[n]$, we will explain how one can amplify speech coming from a certain direction and attenuate noise coming from another. This 'spatial amplification' operation is referred as fixed beamforming, and can be done by either summing delayed or filtered versions of the observed signals. The delays or filters taps depend on the directivity pattern that is sought to achieve, as well as the inherent and physical directivity of the microphones themselves. One problem with fixed beamformers is their high sensitivity to microphone mismatch and imperfection, which is becoming more and more of an issue with the advent of tiny microphones, and especially with densely populated sensor arrays. Another issue at hand is the constraints on the direction of speech and noises that must be known, or at least fixed; the problem of multiple noise sources originating from different directions degrades severely the performance of fixed beamformers.

We will then present more sophisticated methods that vary and adapt the delays and filter coefficients over time, thus referred as adaptive beamforming techniques [8, 65]. Notably, the Generalized Sidelobe Canceller (GSC) combines two fixed beamformers (both for speech and noise) with an Adaptive Noise Canceller (ANC) that seeks to remove remaining residual noise from the extracted speech source. While the GSC requires a priori knowledge of the signal model and sound directions, the MWF (Multichannel Wiener Filter) does not; it exploits both spatial and spectral differences of the speech and noise signals, but inevitably introduces distortion in the speech estimate. The MWF also incurs a very high computational burden on the speech enhancement algorithm [64], making it nearly impractical for real-time mobile applications.

Coherence based filtering methods [38, 51] will then be explained, as well as their origins and details. Its relation to the performance of beamforming techniques will be derived. Finally, we will show how the Mean Squared Coherence (MSC) can be used to construct a speech enhancement post-filter that can later be coupled with conventional single microphone processing techniques.
3.2 Fixed Beamforming

Beamforming in general is illustrated in Figure 3.1.

Taking the Fourier transform of (3.1), the observed signals \( x_i[n] \) can be expressed in the frequency domain as:

\[
X_i(f) = S(f) H_i(f) + N(f)
\]  

(3.4)

The filters \( H_i(f) \) in (3.4) are the transfer functions from a fixed location of the source \( s[n] \) to microphone \( i \). It is more convenient to express the transfer function \( H_i(f) \) as a function of the source's position \( p \), such as:

\[
H_i(f, p)
\]  

(3.5)

Normally, the microphones are assumed to be closely spaced compared to their distance with respect to the source. This is called the far-field assumption, and simplifies analysis while having little or no impact on actual performance. Since the microphone array is far compared to the source, the source’s signal can be considered to arrive parallel to all microphones, as illustrated in Figure 3.2.
The exact position of the source then does not matter; only its angle of incidence $\theta$ with respect to the microphone array does. We can thus rewrite (3.5) as:

$$H_i(f, \theta)$$ (3.6)

The output of the beamformer is then defined by:

$$\hat{S}(f) = G(f)X(f)$$

$$= G(f)[H(f, \theta)S(f) + N(f)]$$ (3.7)

$$= G(f)H(f, \theta)S(f) + G(f)N(f)$$

In the absence of noise, (3.7) can be rewritten as:

$$\hat{S}(f) = G(f)H(f, \theta)S(f)$$

$$= W(f, \theta)S(f)$$ (3.8)

where $W(f, \theta)$ is defined as the directivity pattern of the beamformer. The role of the fixed beamformer can be seen as trying to maximize $W(f, \theta)$ (make it equal to unity) for a particular source direction $\theta$ known a priori.

3.2.1 Microphone characteristics

Microphones are normally classified into two classes, either omnidirectional or directional. Each microphone possesses its own directionality pattern, which gives the attenuation in dB as a function of the angle of incidence of the source for a particular (or all) frequency. This is
most easily represented by means of a polar plot; the most common directionality patterns are illustrated below.

<table>
<thead>
<tr>
<th>Omnidirectional</th>
<th>Cardiod</th>
<th>Hypercardiod</th>
<th>Bi-directional</th>
<th>Shotgun</th>
</tr>
</thead>
</table>

**Figure 3.3 Common microphone directionality patterns**

A differential microphone manifests a cardioid directionality [64]. Entirely omnidirectional microphones with absolutely 0 dB attenuation for all \( \theta \) do not exist in practice, but they are often used in theory to simplify analysis of a microphone array.

Moreover, the microphone itself has a specific frequency response, with a gain that varies as a function of frequency. This microphone directionality and frequency response can be included in the definition of (3.6) as:

\[
H_i(f, \theta) = D_i(f, \theta) \cdot H_i'(f, \theta)
\]  

(3.9)

where \( D_i(f, \theta) \) is characteristic to the microphone itself and \( H_i'(f, \theta) \) is the actual source to microphone transfer function. \( D_i(f, \theta) \) is assumed to be a known characteristic for the microphone in question, but has been shown to change over time [67] due to aging and accumulation of dirt in the acoustic pathway. Nonetheless, the omnidirectional microphone model is often assumed for all microphones \( D_i(f, \theta) \approx D(f) \), and a relatively flat frequency response over the band of interest is often the case \( D(f) \approx 1 \), such that the microphone characteristics can be ignored in (3.9) for that particular case.

### 3.2.2 Delay and sum beamforming

The classical delay and sum beamformer is illustrated in Figure 3.4.
The \textit{endfire linear array} microphone arrangement is used, where microphones are equally spaced, disposed in a linear fashion and assumed to have an omnidirectional response. The microphone outputs $x_i[n]$ are delayed and summed to produce the speech estimate $\hat{s}[n]$ as:

$$\hat{s}[n] = \frac{1}{M} \sum_{i=1}^{M} x_i[n - \tau_i]$$  \hfill (3.10)

The delays $\tau_i$ are chosen as to cancel out the time delays between the arrival of $s[n]$ to each of the microphone (i.e. to resynchronize the signals). As such (see Figure 3.4):

$$\tau_i = -\frac{(d_i - d_1) \cos(\theta)}{c} f_s$$ \hfill (3.11)

where $c$ represents the speed of sound ($340 \text{ m/s}$ is normally used). Assuming an array length $A_i$, the space between each microphone is $d$ and defined as:

$$d = \frac{A_i}{M - 1}$$ \hfill (3.12)

The source is assumed to be at $\theta = 0^\circ$, and thus the directivity pattern $W(f, \theta)$ of the beamformer can be shown to be [64]:

$$W(f, \theta) = \frac{e^{-j2\pi f_1}}{M} \sum_{i=0}^{M-1} e^{j2\pi f_i \left( \frac{d}{c} \cos \theta \right)}.$$ \hfill (3.13)

As such, the magnitude $|W(f, \theta)|$ of the directivity pattern can be derived as [64]:

35
3.2 Fixed Beamforming

\[ |W(f, \theta)| = \frac{1}{M} \frac{\sin(M \varepsilon / 2)}{\sin(\varepsilon / 2)} \]
\[ \varepsilon = 2\pi f \frac{d}{c} (1 - \cos \theta) \]  

(3.14)

Looking at (3.14), it can be seen that \( \sin(M \varepsilon / 2) / \sin(\varepsilon / 2) \approx 1 \) for small \( M \varepsilon / 2 < \pi \). Thus, the delay and sum beamformer provides very little directivity for small \( M \cdot f \cdot d \), or at low frequencies \( (f) \), when the number of microphones is small \( (M) \) and when they are close to each other \( (d) \). Directivity plots of a \( M = 2 \) microphone beamformer at \( d = 2 \) and \( d = 5 \) cm spacing are shown in Figure 3.5.

![Figure 3.5 Directivity of the delay and sum beamformer](image)

(a) \( d = 2 \) cm  
(b) \( d = 5 \) cm

Fixed delay and sum beamformers require larger arrays \( (d = 15 \) cm) to be somewhat directional, especially at low frequencies, and are thus rarely used in portable speech enhancement devices [77].

3.2.3 Filter and sum beamforming

The generalization of the fixed delay and sum beamformer is the filter and sum beamformer, where the delays in (3.10) are replaced by filters. The estimated signal \( \hat{s}[n] \) is then computed as:

36
\[ \hat{s}[n] = \frac{1}{M} \sum_{i=1}^{M} g_i[k] * x_i[n] \]  

(3.15)

The goal is then to design the filters \( g_i[k] \) such as to achieve a desired (most likely highly directional) spatial directivity \( D(f,\theta) \). One way to do that is to minimize a distance measure \( J(f,\theta) \) between the computed directivity \( W(f,\theta) \) of the beamformer and the desired response \( D(f,\theta) \), such as a weighted least square distance:

\[ J_{wls}(f,\theta) = \int \int L(f,\theta) \left[ |W(f,\theta) - D(f,\theta)| \right]^2 df \, d\theta \]  

(3.16)

where \( L(f,\theta) \) is a weighting factor giving more or less importance to certain angles \( \theta \) and frequencies \( f \). Other cost functions and design procedures are discussed and thoroughly explained in [12], as well as the conception of frequency invariant directivity patterns. Nonetheless, adaptive approaches are generally preferred to fixed filter and sum beamformers which require an exhaustive design procedure that is only valid for a specific directivity pattern.

### 3.3 Adaptive Beamforming

#### 3.3.1 Generalized Sidelobe Canceller

The first class of adaptive beamforming algorithms is generally referred as the Generalized Sidelobe Canceller (GSC). Its structure is depicted in Figure 3.6.
3.3 Adaptive Beamforming

The GSC is actually built out of a fixed spatial processor cascaded with an Adaptive Noise Canceller (ANC). In the first stage, the speech reference coming from a known direction $\tilde{s}[n]$ is extracted with a conventional fixed beamformer (see Section 3.2). Next, $M-1$ noise references are created again with a fixed beamformer by steering nulls toward the source in the directivity patterns. As an example [64], the noise references can be created by simply pairwise subtracting the microphone signals $x_i[n]$ that have been time-aligned in the direction of the source. This ensures that the noise contributions $n_i[n]$ in $\tilde{n}_i[n]$ are dominant compared to speech leakage.

Next, a multi-channel adaptive noise canceller removes the leaked noise components in $\tilde{s}[n]$. Ideally, $\tilde{s}[n]$ would not contain noise at all, but the assumptions made about the microphone characteristics, uncertainty of the source position and variable reverberation causes leaks of noise through the speech beamformer. The coefficients $h_i[k]$ are then adapted during noise only periods, with an algorithm such as the Normalized Least Mean Square (NLMS) that approximates the steepest descent; the speech reference is delayed by $\Delta$ samples, as in most
3 MULTI-MICROPHONE SPEECH ENHANCEMENT

LMS implementations, to allow for effectively non-causal $h[k]$ filters. The single channel
NLMS is derived in Appendix 8.5, and can be applied to this multi-channel problem simply
by concatenating the $M-1$ noise references $\bar{n}_i[n]$ and filter coefficients $h[k]$ of the $k^{th}$
window into column vectors. Following the notation of Appendix 8.5, the NLMS adaptation
rule (7.20) can be written as:

$$h_{k+1} = h_k - \mu \frac{\bar{n}_k^T e[n]}{\bar{n}_k^T \bar{n}_k + \varepsilon}$$

(3.17)

where the $e[n]$ can be written as:

$$e[n] = \bar{s}[n] - h_k^T \bar{n}_k$$

(3.18)

Here, $\bar{n}_k$ and $h_k$ are defined as:

$$\bar{n}_k = \begin{bmatrix} \bar{n}_{1,k}^T & \cdots & \bar{n}_{M-1,k}^T \end{bmatrix}^T$$

$$h_k = \begin{bmatrix} h_{1,k}^T & \cdots & h_{M-1,k}^T \end{bmatrix}^T$$

(3.19)

To achieve faster convergence, transform domain LMS or even the RLS algorithm can be
used, at the expense of higher computational expenses.

A theoretical and practical study of the noise reduction limits of the GSC for speech
enhancement was done in [5]. It turns out that its performance is highly dependant on the type
of noise, or conversely on the amount of correlation of noise at the different microphones. We
distinguish generally 3 extreme cases of noise fields:

1. Incoherent: Noise is spatially white. The coherence of noise at any two points in space is
   very low ($\approx 0$).

2. Coherent: Noise is highly directional. The coherence of noise at any two points in space
   is very high ($\approx 1$).
3. **Diffuse**: Noise of equal power propagates in all directions simultaneously. The coherence of noise at any two points is a function only of the distance separating them. This noise field is a good approximation of reverberant environments and of multiple noise source scenarios.

The term *coherence* is loosely used here, but will be later defined in Section 3.4 as a normalized measure of cross-correlation between two signals.

Noise reduction in the GSC thus occurs in two places; in the fixed beamformer (spatial processing) and in the adaptive processing (see Figure 3.6). The study shows that the adaptive processing offers no noise reduction in perfectly incoherent noise fields (i.e. the ANC cannot converge), and infinite noise reduction in perfectly coherent noise fields (i.e. the ANC places a null in the noise direction). Since none of those two scenarios exactly occur in practice, a more interesting case is the diffuse noise field. It is shown that with combinations of $M = \{3 \ldots 7\}$ microphones and $d = \{5, 10, 15, 20\}$ cm spacing, the ANC offers, at best, 1 dB of noise reduction in diffuse noise fields. Noise reduction performance as a function of simulated reverberation time for a white noise source is given in Figure 3.7; a diffuse noise field is said to be loosely equivalent to $>200$ ms reverberation time.

![Graph showing noise reduction performance of the GSC](image)

**Figure 3.7** Noise reduction performance of the GSC
The GSC is thus mainly suited for coherent noise fields, or anechoic environments, where noise is highly directional; as such, its noise reduction capabilities mainly comes from its fixed beamforming processing, which requires precise knowledge of source and noise locations. Consequently, several speech enhancement approaches actually combine the GSC with a traditional Wiener filter or Spectral Subtraction [51] algorithm to remove incoherent noise components.

3.3.2 Multi-channel Wiener Filter

Another appealing approach to adaptive beamforming is the multichannel Wiener filter with unknown reference. Notably, it does not require a priori knowledge of the source or noise directions as in the GSC. Its general structure is illustrated in Figure 3.8.

![Figure 3.8 Structure of the Multichannel Wiener Filter](image)

The MWF minimizes the MSE between the delayed speech contribution in \( x^s[n-\Delta] \), and the sum of the filtered noisy speech observations. Without loss of generality it is usually assumed that \( x^f[n] = x_i[n] \), such that the MWF is estimating the clean speech components of the first microphone, said to be the ‘reference’ microphone. Using matrix notation for the \( k^{th} \) observed window where the signals are placed in column vectors:

\[
x_k = \begin{bmatrix} x_{i,k}^T & \cdots & x_{M,k}^T \end{bmatrix}^T
\]

\[
h_k = \begin{bmatrix} h_{i,k}^T & \cdots & h_{M,k}^T \end{bmatrix}^T
\]

the MWF solves the minimization problem:
3.3 Adaptive Beamforming

\[ h_k = \arg \min_{h_k} E \left[ \left| x^t \left[ n - \Delta \right] - h_k^T x_k \right|^2 \right] \]  \hspace{1cm} (3.21)

The MWF requires a VAD just as the GSC does to estimate the second order statistics of noise (i.e. \( x_k \) during speech pauses). As such, (3.21) is solved in the traditional way by differentiating (3.21) with respect to \( h_k \) and equating to zero. The solution is very similar to the original Wiener derived in Appendix 8.3, and stated in [64] as:

\[ h_k = R^{-1}_{xx} R_{x[n-\delta]} \]  \hspace{1cm} (3.22)

Implementation details of the MWF is beyond the scope of this work, but suffice it to say that direct computation of (3.22) for each sample \( n \) is prohibitively expensive in terms of computational expense. Details of Generalized Singular Value Decomposition (GSVD) and QR Decomposition (QRD) solving approaches are explained in [12] and [59] respectively; these methods are iterative in the sense that the correlation matrices of (3.22) need not be recalculated but rather updated between each frame.

Improvements in SNR are presented in [64] and shown in Figure 3.9 for diffuse noise fields, for endfire arrays of \( M = \{2, 3, 4\} \) microphones and at different frequency array length ratios \( f \cdot A_l \).

![Figure 3.9 SNR Improvements of the MWF](image)

Figure 3.9 SNR Improvements of the MWF

42
For example, to obtain the peak SNR improvement of a $M = 4$ microphone array at $f = 1 \text{ kHz}$ and $f_c = 16 \text{ kHz}$ requires an array length of $A = \frac{(0.03 \text{ m} \cdot 16 \text{ kHz})}{(2 \cdot 1 \text{ kHz})} = 24 \text{ cm}$. The MWF is less sensitive to microphone mismatch than the GSC, but is notably more demanding in terms of computing power; it also requires large size arrays (20 - 30 cm) to achieve appreciable results in voiced frequency bands (1 - 3 kHz).

### 3.4 Coherence based filtering

Another approach at discerning speech from noise is through a measure of the normalized cross-correlation between the microphones [20, 38, 43, 51]. This measure is also known as the coherence, and is often used in the performance analysis of adaptive systems. In parallel to beamforming methods, speech is detected by measuring the coherence, rather than being extracted by additive combination of the input signals; it is assumed that speech is more correlated than noise between the two microphones. As such, a post-filter (to single microphone technique) can be built around this coherence frequency mask to aid in suppressing further residual noise.

#### 3.4.1 Concepts and definitions

The coherence between two signals $x[n]$ and $y[n]$ is defined as in [38]:

$$C_{xy}[k] = \frac{R_{xy}[k]}{\sqrt{R_{xx}[k]R_{yy}[k]}}$$  \hspace{1cm} (3.23)

In the frequency domain, (3.23) can be expressed as:

$$C_{xy}(f) = \frac{P_{xy}(f)}{\sqrt{P_{xx}(f)P_{yy}(f)}}$$  \hspace{1cm} (3.24)

The PSD estimates $P_{xy,k}(f)$ for the $k^{th}$ window are computed in practice by a simple first order IIR filter:

$$P_{xy,k}(f) = \alpha(t_c) \cdot P_{xy,k-1}(f) + (1 - \alpha(t_c)) \cdot X_k^*(f)Y_k^*(f)$$  \hspace{1cm} (3.25)

43
where $\alpha(t_c)$, defined in Appendix 8.4, is close to 1; $t_c = 5\text{ms}$ has proved to offer suitable results, and will be assumed thereof. Since (3.24) is complex valued, a useful scalar measure of coherence is the Magnitude Squared Coherence (MSC), defined as:

$$MSC_{xy}(f) = \left| C_{xy}(f) \right|^2 = \frac{\left| P_{xy}(f) \right|^2}{P_{xx}(f)P_{yy}(f)}$$

(3.26)

To give an insight on the meaning of the coherence function, let us assume that a transfer function exists between $x[n]$ and $y[n]$:

$$y[n] = h[k] * x[n] \leftrightarrow Y(f) = H(f)X(f)$$

(3.27)

If we consider the instantaneous coherence measure $C_{xy}(f)$ and assume for an instant that:

$$P_{xy}(f) = E\{X(f)Y^*(f)\} \approx X(f)Y^*(f)$$

(3.28)

Using (3.28) in (3.24):

$$C_{xy}(f) = \frac{X(f)H(f)^*X(f)^*}{\sqrt{|X(f)|^2|H(f)|^2|X(f)|^2}} = \frac{|X(f)|^2|H(f)|}{|X(f)|^2|H(f)|} e^{-j\angle H(f)}$$

(3.29)

such that:

$$MSC_{xy}(f) = \left| e^{-j\angle H(f)} \right|^2 = 1$$

(3.30)

That is, the MSC between $x[n]$ and $y[n]$ is (theoretically) equal to unity when a transfer function exists between them. Similarly, one can prove that $MSC_{xy}(f) = 0$ when no transfer function exists between the two signals ($X(f)$ and $Y(f)^*$ are then uncorrelated, $P_{xy}(f) = 0$ and so (3.26) = 0).

From Section 3.3.1 and the previous discussion, it should now be obvious why a directional noise source exhibits high coherence (a single transfer function exists between the two
microphones for that particular direction), while a diffuse noise source exhibits lower coherence (multiple transfer functions exist, in all directions, between the two microphones).

3.4.2 Connection with ANC performance

As explained before, the performance of the ANC presented in the GSC is proportional to the coherence between the adapted signals. Calculating the PSD of the ANC error $e[n]$ (3.18), $P_{ee}(f)$ equals:

$$P_{ee}(f) = \left| H(f) - \frac{P_{sh}(f)}{P_{sh}(f)} \right|^2 P_{sh}(f) + (1 - C_{sh}(f)) P_{sh}(f) \quad (3.31)$$

The proof of (3.31) is derived Appendix 8.6, by substituting $x[n] \rightarrow \tilde{s}[n]$ and $y[n] \rightarrow \tilde{n}[n]$. Assuming convergence of the ANC (by the NLMS algorithm for example), $H(f)$ converges to $P_{sh}(f) / P_{sh}(f)$ (i.e. the Wiener filter) and $P_{ee}(f)$ is proportional to one minus the coherence $C_{sh}(f)$ between the speech and noise estimated by the fixed beamformers. In turn, this coherence will be maximal when the residual noise in $\tilde{s}[n]$ is related by a transfer function $g[n]$ to $\tilde{n}[n]$, that is when:

$$\tilde{s}[n] = \tilde{s}[n] + g[n] * \tilde{n}[n] \quad (3.32)$$

In this case when $C_{sh}(f)$ is maximized, $P_{ee}(f)$ is minimized (see (3.31)) which justifies the higher performances of the GSC in directional noise scenarios (and vice-versa).

3.4.3 Post-filtering

It has thus been suggested to use the coherence measure between two channels to isolate speech from noise, when the former is assumed to be more correlated than the later. The assumed observation model is the two microphone case of eq. (3.1), with $x_i[n]$ as the reference microphone and $x_s[n]$ as the secondary microphone. Under this model, we will first study the theoretical coherence measure of diffuse noise, a typical and realistic noise scenario,
and we will see how the coherence measure can be used to build a suitable speech enhancement post-filter.

3.4.3.1 Diffuse noise coherence

Noise often comes from different directions and is also often produced in reverberant environments; babble noise, office noise and car engine noise are such examples. A suitable model in that case is the diffuse noise scenario, by which definition noise originates equally likely in all 3 directions. A formal derivation of the coherence function for diffuse noise was done in [23] and proceeds by evaluating the coherence function between two microphones for a source coming in a direction $\theta$ (such as illustrated in Figure 3.4), and then integrating over the whole possible $\theta$ and $\phi$ spheroid coordinates. After simplification, they arrive at the following expression:

$$ C_\nu(f) = \frac{\sin(2\pi f \cdot d / c)}{2\pi f \cdot d / c} = \text{sinc}(2\pi fd / c) \quad (3.33) $$

The theoretical coherence is thus real-valued and only depends on the distance $d$ separating the two microphones.

To verify (3.33), an actual measure of coherence was made on noise only segments of a multimicrophone speech database. The database in question was the CMU Microphone Array database [66] where the first 2 of the 8 microphones were used to compute the coherence using the PSD approximations of (3.31). The microphones were separated by 7 cm and the samples were collected in a noisy computer lab, with many power supply and disk drive fans. A plot of the mean real part of the measured coherence overlaid with the theoretical coherence given by (3.33) is presented in Figure 3.10.
Figure 3.10 Theoretical and measured diffuse noise coherence

As can be seen, the observed coherence is found to coarsely match its theoretical prediction. Diffuse noise coherence being higher in the lower frequency bands, care will be required to avoid confusing noise for coherent speech; using only low noise-coherent subbands to estimate speech coherence was done in [39, 52] and was shown to reduce speech distortion.

3.4.3.2 Proposed filters

Exploiting speech coherence between the microphones, Lebouquin [38] suggested several ways to use that measure for the purpose of noise reduction. The first method is to weight the in-phase frequency bands with the magnitude of the coherence. This is done as:

\[
\hat{S}(f) = \left[ \frac{X_1(f) + A(f)X_2(f)}{2} \right] |C_{\alpha\beta}(f)|
\]

where

\[
A(f) = \frac{X_1(f)X_2^*(f)}{|X_1(f)||X_2(f)|}
\]

The magnitude of the coherence is thus applied as a weighting factor. Another technique suggests using thresholds on the coherence to filter differently coherent and non-coherent bands.
3.4 COHERENCE BASED FILTERING

\[
\hat{S}(f) = \begin{cases} 
C_{xy}(f) < C_{\min} & C_{\min}^{\alpha} X(f) \\
otherwise & |C_{xy}(f)|^{\alpha} X(f) \\
C_{xy}(f) > C_{\max} & X(f)
\end{cases} 
\quad (3.36)
\]

where \( \alpha > 1 \). Another gain function was suggested [20, 73] and makes uses of the MSC instead of the coherence itself:

\[
H_{MSC}(f) = e^{\alpha - \alpha/MSC_{xy}(f)} 
\quad (3.37)
\]

It can also be parameterized through \( \alpha \) to control the aggressiveness of the noise reduction; a plot of \( H_{MSC}(f) \) for \( \alpha = [1, 2, 3] \) is shown in Figure 3.11.

![Figure 3.11 H_{MSC} as a function of the MSC](image)

An easier way to visualize \( H_{MSC}(f) \) is through its attenuation as a function of the input SNRs of \( X_1(f) \) and \( X_2(f) \). Defining temporarily the SNR at a microphone as the ratio of the speech components (not the original speech) to the noise components as:

\[
SNR_i(f) = \frac{|H_i(f)|^2 P_{s,i}(f)}{P_{n,i}(f)} 
\quad (3.38)
\]

we can rewrite the definition of the MSC in (3.26) as:
\[ MSC_{xy}(f) = \frac{\text{SNR}_1(f) \cdot \text{SNR}_2(f)}{(1 + \text{SNR}_1(f))(1 + \text{SNR}_2(f))} \]  
\[ (3.39) \]

Similarly as was done for the noise suppression rules (Section 2.7.1), the gain function \( H_{MSC}(f) \) in dB has been plotted in Figure 3.12 for equal input SNRs at both microphones.

![Figure 3.12 H_{MSC} in dB as a function of the input SNRs](image)

The parameter \( \alpha \) can be chosen as a function of the difference in speech energy at both microphones, depending on the recording setup. \( H_{MSC}(f) \) has been used in [20] and was shown to subjectively provide further noise suppression when used as a post-filter to single microphone techniques.

### 3.5 Other Methods

To complete this overview of multi-microphone techniques, we briefly present two other lesser known methods that are based neither on beamforming or coherence measurements. Blind source separation and phase-error approaches are summarized, and references to particular in-depth publications on the subjects are given; this section aims to convince the reader that numerous other fields of research can be applied to our speech enhancement problem.
3.5 Other Methods

3.5.1 Blind Source Separation

Attempts at blind speech denoising using Independent Component Analysis (ICA) have been made [57, 75], but the results have been so far limited to small improvements on word recognition engines. ICA based algorithms are not exactly well suited to blind speech denoising. Restrictions on the non signal gaussianity, permutation problems as well as a priori knowledge requirement on the number of sources impose hard restrictions on noise scenarios which are most often violated in real environments.

Moreover, blind source separation algorithms assume instantaneous mixing of sources, which is obviously not the case in reverberant environments. Solutions have been proposed for convolutive mixtures by transforming the time domain problem to a series of instantaneous ICA problems in the frequency domain [60], but inter-window permutation ambiguities complicate the procedure. This has been partially solved by analyzing directivity patterns formed by the separation matrices, relying on the fact that the sources are discretely located in space, an obviously false assumption for highly diffuse noise fields.

For all those reasons, and because of its high complexity, ICA based denoising is not yet considered on par in terms of usability and performance with its spectral domain algorithmic counterparts.

3.5.2 Phase-Error based Filtering

Another recent approach to speech enhancement consists in estimating a bound on the SNR of a particular frequency band by observing the phase difference between pairs of Fourier transform coefficients [1, 25, 62]. Assuming a bi-microphone observation model without reverberation (results for reverberant environments are nonetheless presented), we can write (as a specific case of (3.1)):

\[
\begin{align*}
    x_1[n] &= s[n] + n_1[n] \\
    x_2[n] &= s[n-\tau] + n_2[n]
\end{align*}
\]  

(3.40)
Assuming that the amplitude of the noise spectrum is similar in both channels, that is \(|N_{1,k}(f)| \approx |N_{2,k}(f)| = N_k(f)|\), the Fourier transform of (3.40) for the \(k\)th data frame can be written as:

\[
\begin{align*}
X_{1,k}(f) &= |S_k(f)| e^{j\varphi_k(f)} + |N_k(f)| e^{j\varphi_N(f)} \\
X_{2,k}(f) &= |S_k(f)| e^{j\varphi_k(f) - j2\pi f \tau} + |N_k(f)| e^{j\varphi_N(f)}
\end{align*}
\]  

(3.41)

The inter microphone Time Delay of Arrival (TDOA) \(\tau\) is assumed to be known a priori, but can be estimated using techniques such as the Phase Transform (PHAT) cross-correlation method used in source localization problems [21]. The idea is then to define a measure called the phase error \(\theta_k(f)\) defined as:

\[
\theta_k(f) = \angle X_{1,k}(f) - \angle X_{2,k}(f) - 2\pi f \tau
\]  

(3.42)

Observing (3.41) and (3.42) it can be seen that \(\theta_k(f)\) is 0 in the speech only case (i.e. if \(|N_k(f)| = 0\)), and non-zero in the presence of noise; it thus intuitively suggests a measure of the signal to noise ratio. This fact was recognized by the authors, who derived an SNR bound as a function of \(\theta_k(f)\) [1]:

\[
\text{SNR}_k(f) = \frac{|S_k(f)|^2}{|N_k(f)|^2} \leq \frac{1}{\sin^2(\theta_k(f)/2)} \]  

(3.43)

Conceptually, the method is simply a different (multi-microphone) technique to estimate the SNR, parallel to the noise estimation techniques presented in Section 2.7.2. As such, it can be used with previously presented suppression rules such as the Wiener filter. Using (3.43) in (2.22) the spectral gain \(H_k(w)\) can be expressed as:

\[
H_k(f) \leq \frac{1}{1 + \sin^2(\theta_k(f)/2)}
\]  

(3.44)

The authors suggest using an approximation of (3.44) with a factor \(\gamma\) that controls the aggressiveness of the spectral mask:
3.6 Summary

Multi-microphone speech enhancement techniques were presented in this chapter. Fixed and adaptive beamforming techniques were explained and were shown to offer especially good noise reduction when the noise sources are directional and microphone spacing is high. The concept of coherence was then introduced and its relation to noise directionality was detailed; the particular case of diffuse noise sources was studied and seen to be an appropriate model of real life recordings. Post-filtering functions based on the MSC were presented, as well as their gain curves as a function of the input SNRs. Blind source separation and phase error based filtering were finally overviewed to wrap-up the subject.
CHAPTER 4
PERCEPTUAL SPEECH ENHANCEMENT

Research in auditory perception has recently spurred the development of perceptual models, otherwise known as masking models, which aim to emulate the human ear and thus represent the perceived spectral information of an audio signal. The field of lossy audio coding is particularly known for pushing the boundaries and aggressively using those masking models in view of distributing coding bits proportional to their perceptual relevance. A thorough and comprehensive review of perceptual audio coding techniques was published in [55] and is given as a reference for an introduction on the subject.

Perceptual models have been recently applied to the field of speech enhancement, in which the general idea is to subtract perceived noise only. Without going into minute details, we will present a typical and popular perceptual model, the Perceptual Evaluation of Audio Quality (PEAQ) masking model, and we will explain the most popular methods in which it can be used to enhance speech buried in noise. Since the primary object of this thesis is not about perceptual noise suppression, the concepts will be covered briefly, yet concisely; the goal is to give the reader an overview of the perceptual information that can be exploited to reduce speech distortion resulting from traditional spectral processing.

4.1 The PEAQ masking model

The ultimate goal of the masking model is to compute a masking threshold $M_X[m]$ from a spectrum $X[m]$. The PEAQ perceptual model is illustrated in Figure 4.1.

Figure 4.1 PEAQ masking model
4.1 The PEAQ Masking Model

The details of each block is explained in [69], along with formulas and values required for practical implementation. The procedure can be summarized as follows:

1. **Ear transfer function:** \( |X[m]| \) is multiplied by a weighting function that accounts for the frequency dependent sensitivity of the human hear; interestingly enough, this function peaks between 2 kHz and 4 kHz where most of speech’s energy is concentrated. A weighted spectrum \( X_w[m] \) is thus obtained.

2. **Bark domain conversion:** The weighted power spectrum \( |X_w[m]|^2 \) bins are grouped in 0.25 Bark bands as \( X_b[b] \). The Bark scale is logarithmic in nature and the bandwidths of the bands thus increase as a function of center frequency.

3. **Adding of internal noise:** A frequency dependent internal noise floor \( N[b] \) that accounts for the absolute threshold of hearing (an internal noise floor naturally produced by the human body) is added to \( X_b[b] \); \( X_{bn}[b] \) is obtained.

4. **Critical band spreading:** The critical bands in \( X_{bn}[b] \) are convolved with an amplitude dependent triangular spreading function \( S[b] \) having a right hand slope that decreases with increasing amplitude (higher amplitude incurs more masking); \( X_{bns}[b] \) results from that convolution. Operating in the Bark domain simplifies the design of the spreading function.

5. **Temporal spreading:** The spread energies \( X_{bns}[b] \) are smoothed in time with a frequency dependent IIR filter to account for forward and backward temporal masking; the mask pattern \( X_{bnsf}[b] \) is obtained.
6. **Frequency domain conversion**: Finally, the mask pattern $X_{\text{BTCR}}[b]$ is converted back into the frequency domain through linear or spline interpolation; the masking threshold $M_X[m]$ is obtained.

An example of the PEAQ masking threshold for a voiced speech spectrum is illustrated in Figure 4.2.

![Graph showing PEAQ masking threshold](image)

**Figure 4.2 Sample PEAQ masking threshold**

We recall that the implementation details of those operations are well documented in [69], as well as in [34]. For the purpose of our research, the default PEAQ implementation had to be modified and generalized to work at other sampling frequencies than the default 8 kHz.

### 4.2 Application to speech enhancement

#### 4.2.1 Tsoukala & Soulodre's method

In order to obtain a signal with the same psychoacoustic representation as the clean signal, Tsoukalas et al [70] state that the ideal filter should be (similarly to the Wiener filter, see eq. (2.16)):

$$H_T(f) = \frac{M_s(f)}{M_X(f)}$$  \hspace{1cm} (4.1)
4.2 APPLICATION TO SPEECH ENHANCEMENT

where the previously presented PEAQ model can be used to calculate the masking threshold. Since \( P_{a}(f) \) is obviously unknown, it can be approximated as \( \hat{P}_{ss}(f) = P_{xx}(f) - \hat{P}_{nx}(f) \) (by estimating the noise spectrum, see Section 2.7.2) such that:

\[
H_{\tau}(f) = \frac{M_{x}(f) - M_{\beta}(f)}{M_{x}(f)} = 1 - \frac{M_{\beta}(f)}{M_{x}(f)}
\]  

(4.2)

Soulodre points out in [63] that masking varies with amplitude level and is highly non-linear. Eq. (4.2) shall thus be rewritten as:

\[
H_{\tau}(f) = \frac{M_{x-\beta}(f)}{M_{x}(f)}
\]

(4.3)

This has been used successfully in speech and audio restoration of signals corrupted by camera noise [63].

4.2.2 Virag’s method

Virag devised a method to control the oversubtraction and noise floor factors \( \alpha \) and \( \beta \) of Berouti’s GSS rule (see eq. (2.35)) using the masking threshold \( M_{x}(f) \) calculated from clean speech estimated using standard spectral subtraction. The idea is to subtract less in highly masked bands (higher \( M_{x}(f) \), lower \( \alpha \)), and to subtract more in poorly masked bands (lower \( M_{x}(f) \), higher \( \alpha \)). This can be formulated as:

\[
\alpha(f) = \begin{cases} 
\alpha_{\min} & M_{x}(f) = M_{\max} \\
\text{otherwise} & F_{\alpha} \left[ \alpha_{\min}, \alpha_{\max}, M_{x}(f) \right] \\
\alpha_{\max} & M_{x}(f) = M_{\min}
\end{cases}
\]

(4.4)

where \( F_{\alpha} \left[ \alpha_{\min}, \alpha_{\max}, M_{x}(f) \right] \) is a linear function that maps \( M_{x}(f) \) to \( \alpha \) between \( \alpha_{\min} \) and \( \alpha_{\max} \) as explained above. The same logic follows for \( \beta \), whereas a highly masked band should adopt a lower noise floor (higher \( M_{x}(f) \), lower \( \beta \)) and vice versa. Comparison to standard GSS suggests 2 - 3 dB improvements in segmental SNR and slight preference in subjective MOS testing [74].
4 PERCEPTUAL SPEECH ENHANCEMENT

4.2.3 Musical noise suppression

Another use of perceptual models has been suggested where a direct attempt is made at
detecting and removing perceived tonal components in the output masking threshold \( M_s(f) \)
not present in the input masking threshold \( M_x(f) \) [30]. These are assumed to be musical
noise artefacts (see Section 2.7.3.1) well known to be introduced by spectral suppression
rules, and are to be attenuated just under the masking threshold of the enhanced speech
spectrum.

Special care is taken to avoid false detection of speech as musical noise (by imposing
bandwidth constraints), and spectrum smoothing is performed on neighbouring spectral bins.
The post-filter is combined with traditional Wiener filtering; objective improvements in
perceptual spectral distance measures are presented, but only with additive white Gaussian
noise. No subjective results are presented on real noisy samples.

4.3 Summary

We briefly presented ways to incorporate perceptual elements in speech enhancement
algorithms. The PEAQ perceptual model was presented as a classic, yet representative
example of masking threshold models, and its constituting blocks were briefly summarized in
a comprehensive manner. Ways to exploit this model were shown, first by performing
‘perceptual’ spectral enhancement, then by complementing general spectral subtraction rules,
and finally by directly attacking the problem of musical noise. Objective improvements are
generally presented in related research, but demonstration of subjective superiority over state
of the art non-perceptual algorithms using real noisy samples has yet to be performed.
CHAPTER 5
PROPOSED SYSTEM

5.1 Overview

The system we propose aims to combine state of the art single microphone processing techniques with a suitable multi-microphone post-filter. Since spectral enhancement methods are better known to provide significant subjective improvements in quality compared to other approaches, they were chosen both for the single and multi-microphone processing part of our system. The system is illustrated in Figure 5.1.

![System Overview Diagram]

Figure 5.1 Proposed system overview

After experimentation, the choice was made not to include perceptual speech enhancement methods in our system. Multi-channel perceptual speech enhancement goes beyond the scope of this thesis, and its definite subjective improvements on real noisy speech over non-perceptual techniques haven’t been established so far. Effort was rather put in choosing, implementing and experimenting with robust single and multi-microphone methods presented in this work, as well as ensuring good subjective performance on typical real life noise scenarios.

We will thus detail, in the next two sections, the insides of our single microphone processing box and then our multi-microphone post-filter (Figure 5.1).

5.2 Single microphone speech enhancement

Our single microphone processing is divided into three main blocks:

1. Noise estimation
2. SNR estimation
3. Filtering with a suppression rule

The system is illustrated in details in Figure 5.2.

---

**Figure 5.2 Proposed single microphone speech enhancement**

MCRA noise estimation was chosen for its robustness in varying noise environments; since it inherently contains a per-band soft-decision VAD, it can track noise simultaneously in different voiced and unvoiced part of the spectrum. The decision directed approach was used to estimate $SNR_{prio}$ as it provides good results and is the most widely used technique in the literature. The four suppression rules $H_{w}$, $H_{SS}$, $H_{EMSR}$ and $H_{log_{EMSR}}$ were then chosen to be compared, as they offer significant noise suppression in low input SNRs (see Figure 2.7). To improve the performance of $H_{w}$ and $H_{SS}$, $SNR_{prio}$ was used instead of $SNR_{post}$ as it is known...
that it is the SNR smoothing of $SNR_{pri}$ in the EMSR suppression rules that virtually eliminates the appearance of musical noise [7]; significant subjective improvements were informally confirmed by using this simple substitution.

### 5.3 Multi-microphone post-filter

Our multi-microphone post-filter had to be based either on beamforming techniques (fixed or adaptive) or coherence-based filtering. Experiments on real life recorded noisy samples showed little convergence of adaptive beamformers in noise only periods; listening at the audio samples indicated the presence of multiple or reverberant noise sources in most cases. Measuring the noise coherence as was done in Section 3.4.3.1 further confirmed that a diffuse noise scenario best represents the environments commonly encountered.

For those reasons, we chose to adopt a post-filter based on speech coherence instead of beamforming, which would otherwise also require knowledge of the speaker's position. As mentioned in Section 3.4.3.1, care must be taken not to confuse high noise coherence in the low frequencies for speech presence; weighting in the log domain of the post-filter by the complement of the theoretical diffuse noise field coherence thus minimizes speech distortion in the low-frequency bands. The system is illustrated in Figure 5.3.
The weighting function is designed such that when the theoretical noise field coherence is close to unity, the effect of the post-filter is cancelled and vice versa (see Figure 3.10). An effective way to achieve that function is to multiply the post-filter $G_{MSC}$ in the log domain (exponentiation in the linear domain) by one minus the theoretical noise field coherence (eq. (3.33)). This is formulated as:

$$W(f) = \alpha \left(1 - \text{sinc} \left(\frac{2\pi fd}{c}\right)\right)$$

(5.1)

With discrete frequency bins, this becomes:

$$W[m] = \alpha \left(1 - \text{sinc} \left(\frac{2\pi md}{Nc}\right)\right)$$

(5.2)

A higher $\alpha$ guarantees less speech distortion at the expense of more residual noise; $\alpha = 1$ was used throughout our experiments.
One variation of the post-filter was presented in [20] where the MSC and the post-filter were calculated in quarter Bark subbands (similar to the PEAQ model, see Section 4.1) and later interpolated back in the linear frequency domain such as to uniformly process perceptually similar frequencies. While results showed slight subjective preference to this 'perceptual' smoothing of the post-filter, we will hold to its non-smoothed version for implementation and computational simplicity, as well as result reproducibility.
CHAPTER 6
EXPERIMENTS & RESULTS

In order to evaluate the performance of the system, both objective and subjective experiments were carried out. Objective evaluation was conducted to compare the different single microphone algorithms, as well as to quantify the improvement provided by the multi-microphone post-filter. Subjective evaluation involved listening tests and aimed to determine perceptual preference of the multi microphone post-filter over single microphone enhancement only. Keeping in mind that limited resources (time and listeners) were available for subjective testing, a constrained yet representative set of audio samples was used to gather the results; the Matlab code, samples and documentation is thus provided with this research thesis [19] to let users experiment with their own samples.

6.1 Objective evaluation

Two different experiments were conducted to objectively compare both the single and the bi-microphone speech enhancement systems. The two evaluations aimed respectively to:

1. determine the best single-channel speech suppression rule

2. using that best single-channel suppression rule, determine the effect of the bi-microphone post-filter.

Objective evaluation essentially consists in measuring a distance between the enhanced speech (the output) and the original clean speech (the input). A number of distance measures have been proposed for speech processing algorithms [33], from which we will use the popular Segmental SNR Improvement $SEGSR_{\text{IMPR},db}$ to measure the reduction in residual noise, and the Linear Prediction Log-Spectral Distortion $LPLSD_{db}$ to measure the amount of resulting speech distortion; these distance measures are defined and derived in Appendix 8.7. The system is illustrated in Figure 6.1.
Since the original monophonic clean speech sample $s[n]$ had to be known in order to compute a distance measure, stereophonic clean speech had to be simulated, according to typical recording conditions. For that matter, we used the Room Image Method [2], with a room of $5 \times 5 \times 3$ meters, a reflection coefficient of 0.2 and $n = 2$ levels of reflection; the microphones were simulated 8 cm apart, 30 cm away from the source and in the center of the room. White Gaussian noise was injected at different input SNRs and the $\text{LPLSD}_{\text{dB}}$ (see eq. (7.31)) was computed between the enhanced and original speech version; the idea was to determine which suppression rule, in the initial single channel processing, yields the lowest speech distortion. The results are plotted in Figure 6.2.
Figure 6.2 LPLSD of single microphone suppression rules

The Ephraim & Malah $H_{EMSR}$ and $H_{logEMSR}$ rules were shown to produce the least amount of perceivable speech distortion, especially under increasing values of input SNRs. The $H_{logEMSR}$ rule was thus chosen as the 'best' suppression rule, and used in the second experiment were the effects of the post-filter were studied. The improvement in segmental SNR was then computed and is shown in Figure 6.3, with and without the post-filter.

Figure 6.3 SEGSRImprdb effects of the post-filter
As can be seen, up to 10 dB $\text{SEGSNR}_{db}$ improvements can be observed in low input SNR conditions with the use of the post-filter; its effect lessens and becomes negligible as the input SNR increases ($< 10$ dB). A small ($< 0.5$ dB) but consistent increase in $\text{LPLSD}_{db}$ was also observed over the range of input SNRs using the post-filter; this indicates that the post-filter reduces residual noise at the expense of a slight increase in speech distortion. The amount of speech distortion can be reduced by tuning the aggressiveness of the post-filter through $\alpha$ (see Figure 3.12), at the expense of less residual noise reduction; its effect were seen to be nonetheless globally beneficial under normal ($< 10$ dB) noisy input SNR conditions.

6.2 Subjective evaluation

Subjective testing was carried out to determine the preference (or not) of samples processed with and without the bi-microphone post-filter. Three (3) samples were taken from the CMU Microphone Array database [66], with speech recorded in a noisy computer lab (ambient disk drive and fan noise) at 16 kHz and 16 bits per sample. The WD-063 Panasonic omnidirectional microphones were spaced 7 cm apart and roughly one meter away from the speaker. Since the clean speech and noise signals were unavailable, the exact input SNR could not be determined, but was estimated to be between 5 and 10dB; the noise was more low-pass in nature and more or less constant in energy over the 6 seconds speech samples.

Spectrograms of one of the samples taken at the input (noisy speech), after the single microphone processing, and after the bi-microphone post-filter are shown in Figure 6.4.

![Spectrograms](image)

| Noisy speech | Single microphone processing only | Single microphone + bi-microphone post-filter |

Figure 6.4 Spectrograms of a sample processed using the proposed system
Twelve (12) subjects were presented with 12 pairs of samples processed with and without the coherence based post-filter; the 12 pairs were generated by processing the three database samples with the four single channel suppression rules evaluated above (see Figure 5.2). They were asked to give an ‘overall speech quality’ preference of either sample or no preference at all. The results are shown in Table 6.1.

<table>
<thead>
<tr>
<th>Suppression rule</th>
<th>without post-filter</th>
<th>with post-filter</th>
<th>No preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{EMSR}$</td>
<td>33%</td>
<td>67%</td>
<td>0%</td>
</tr>
<tr>
<td>$H_{logEMA}$</td>
<td>39%</td>
<td>58%</td>
<td>3%</td>
</tr>
<tr>
<td>$H_w$</td>
<td>19%</td>
<td>42%</td>
<td>39%</td>
</tr>
<tr>
<td>$H_{ss}$</td>
<td>33%</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>Overall</td>
<td>31%</td>
<td>50%</td>
<td>19%</td>
</tr>
</tbody>
</table>

Table 6.1 Subjective post-filter results

It can be seen that the post-filter has a net positive effect on the EMSR suppression rules, for which the single microphone enhancement introduced less speech distortion at the expense of more residual noise. Comments from the subjects indicated cleaner speech from the post-filtered samples with a reduced reverberant effect (produced by single microphone EMSR rules). The preference is less obvious for the Wiener and Generalized Spectral Subtraction rules; this is explained by the fact that those two rules already aggressively reduce noise to such an extent that the post-filter can do little to further enhance the speech estimate.

Even though the overall preference of the post-filter scored 50%, it was found to be improving on single channel enhancement in the perceptually preferred EMSR rules. Additional subjective testing with those two rules on more samples recorded at different input SNRs would be required to further assess the performance of the post-filter; since subjective evaluations are costly both in time and effort from the (free willing) listeners, they could not be conducted. Testing at 8 kHz narrowband sampling rates could also give insight on different subjective results since speech distortion is already high (most of the high frequency vocal
cues are already gone), and the further reduction in residual noise might outweigh even more the small distortion incurred using the post-filter.

Moreover, since few researches involving subjective evaluation of speech enhancement algorithms exist, the actual criteria from which listeners should be judging samples has yet to be established. In his comparative study of speech enhancement algorithms, Hu [27] noted that listeners repeatedly score the original noisy speech higher than the processed versions when asked about the overall quality of the signals; however, asking independently about residual noise intrusiveness and speech quality helps to isolate which of the two factors affects the listening experience the most.

6.3 Performance and experimental setup

Several factors will affect the real-life performance of our multi-microphone post-filter. Even though a single setup was used to perform the experiments (two microphones spaced 7 cm apart and 30 cm away from the speaker), we can use the results derived in Section 3.4 to infer on the effects of changing the sampling environment.

The spacing of the microphones will directly affect the estimated $M \text{SC}_{xy}(f)$ used in our post-filter. Still assuming a non-incoherent noise field, bringing the microphones closer will increase the noise coherence between them; that will lead to an over-estimated $M \text{SC}_{xy}(f)$ during noise-only frequency bands, thus wrongly identifying noise as coherent speech, yielding a noisier filtered output.

On the other hand, spacing the microphones further apart too much will decrease the speech coherence between them, especially in reverberant environments, thus lowering the estimated $M \text{SC}_{xy}(f)$ in speech frequency bands. The speech will then be confused for noise which will yield a distorted output where speech components have been too much attenuated; we are facing the common noise trade-off versus speech distortion dilemma present in most speech enhancement systems.
Also, positioning the microphones too far from the speaker will simply decrease both input SNRs; on the other hand, putting them too close might just be physically impossible, or could severely interfere with their omnidirectional directivity assumption flat across the whole spectrum (most electret microphones actually have a cradiod-like directivity pattern).

Another factor that could affect the performance of our post-filter is the assumed theoretical diffuse noise field coherence; scenarios where directional noise sources tend to dominate will affect negatively the performance of our system. Passing cars or localized interfering speakers are examples of directional noise sources that are harder to cancel out using the proposed technique. However, as corroborated by the measurements in Section 3.4.3.1, a diffuse noise field is often encountered and is thus a valid assumption in most ambient noise scenarios.

Again, time and resource constraints limited the scope of our subjective evaluations; further experiments could yield limiting numbers on microphone and speaker positioning, giving insights on the bounds up to which the proposed system can function adequately.
CHAPTER 7
CONCLUSION

In this research, the problem of acoustic noise suppression was explored and several solutions were proposed. Several single channel noise suppression techniques were implemented and compared; the EMSR rules using MCRA noise estimation were shown to significantly enhance overall speech quality in noisy environments while minimizing speech distortion. The concepts of noise field coherence and noise directivity were explained, as well as to why beamforming techniques are more suited towards directional noise interferers. Analysis of real world noisy samples showed a higher correlation to a diffuse noise scenario. As a result, a novel coherence-based bi-microphone post-filter was suggested, as well as its corresponding weighting function dependant on the theoretical noise field aimed at avoiding speech distortion. Using real-word samples, objective and subjective results were presented and the post-filter was shown to improve on the performance of state of the art single microphone noise suppression systems.

7.1 Summary

The aim of this thesis was to present, before and after all, a concise summary of the field of acoustical noise suppression, while suggesting a small improvement on the best existing techniques; focus was kept on using a writing style that inherently describes ways to implement the suggested methods, while showing their limits and interrelation to each other.

In Chapter 1, different contexts justifying the need for speech enhancement algorithms were introduced, such as mobile telephony, hearing aids and in-vehicle communication systems. The different solutions to the problem were coarsely classified into three broad categories that constitute the state of the art review of this work; single microphone, multi-microphone and perceptual speech enhancement methods. This research contributions to the field were then summarized, as well as a more detailed organisation of the document.

Single microphone speech enhancement methods were reviewed in Chapter 2. The short time analysis and synthesis framework was presented, as well as the effects of the different windows available. Several approaches such as wavelet, subspace and auto-regressive speech
modelling methods were detailed but shown to exhibit significantly more distortions than the most popular spectral modeling techniques. The most popular spectral domain suppression rules were then introduced, with emphasis on how to implement them and how they were derived; those rules were then compared against each other for different noise input levels. A recent noise estimation technique, required for the implementation of the previously discussed rules, was summarized; MCRA noise estimation was shown to consistently track the underlying noise spectrum even during speech periods. Some artefacts to spectral processing were revealed, notably musical noise, as well as a justification on why the EMSR rules minimize those particular artefacts.

Chapter 3 extended the solution space to multi-microphone methods. It was explained on how selective attenuation of sounds coming from predefined directions can be achieved using fixed beamforming, and how that can be applied to speech enhancement. The cases where the noise is either changing or its direction unknown were discussed, as well as how adaptive beamforming can be used in such situations. Coherence was then introduced first to explain why beamforming methods are most suited for directional noise sources, and then to present coherence-based post-filtering techniques. Real-world noise was shown to be more diffuse (ie, reverberant) than directional, and as such justifies the preference to coherence adapted techniques as opposed to beamforming approaches. Other lesser used methods were finally briefly covered to wrap-up the topic.

Ways to incorporate perceptual criteria into single microphone enhancement methods were then shown in Chapter 4. First, the PEAQ masking model was briefly dissected and its constituting parts analyzed. Then, its integration in perceptually based spectral suppression rules were covered, as well as direct attempts at suppressing musical noise. References indicating subjective improvements on synthetically generated noisy speech samples were then given, even though lack of significant results using real-world samples was also mentioned.

In Chapter 4 we proposed a novel system combining MCRA noise estimation and several noise suppression rules for single microphone processing, supplemented by a coherence-based
bi-microphone post-filter. The post-filter was explained to be logarithmically weighted by a function complementary to the theoretical diffuse noise field, to avoid confusing speech coherence with noise coherence; this aims to reduce the inevitable speech distortion while maximizing the amount of noise suppressed.

Subjective and objective results of the proposed system were provided in Chapter 5. Objectively comparing suppression rules first revealed lower spectral distortion measures using the EMSR suppression rules. Choosing those rules as the baseline for our single microphone processing, the bi-microphone post-filter was added to the system; while a slight increase in speech distortion was observed, significant improvements in segmental SNR were also noted. Subjective evaluation using listening tests confirmed the overall preference of using the post-filter over not using it at all.

7.2 Further research directions

Several problems and issues remain unanswered in the field of acoustical noise suppression. Even though we have shown that a bi-microphone post-filter can improve on state of the art methods, speech is not perfectly recovered and the performance of the algorithm might vary under different, yet untested noise scenarios. We will thus suggest various research avenues that could be investigated to develop an even better bi-microphone noise suppression system, or at least ease its integration on mobile devices, such that it can be applied to a broader range of noisy speech settings.

7.2.1 Alternative microphone placement

Since the goal was to handle diffuse noise scenarios without much assumption about speaker localisation, the system was tested on samples using a fairly standard setting where the speaker faces two closely spaced omnidirectional microphones. Having a vague idea of the relative position of the speaker to the microphones, attempts could be made at using somewhat directional microphones (differential microphones, for example) to at least isolate sounds coming from the speaker’s direction. Extension to our coherence based post-filter could be also made using more than two microphones, where all possible pairs of microphones could be taken to evaluate the coherence measures used to construct the post-filter. That would
further raise the question of whether equally spacing the microphones linearly is the optimum configuration for such an application.

7.2.2 Hybrid method for different noise scenarios
The proposed post-filter was optimized for diffuse, reverberant noise scenarios, since this is a common case of speech degraded by ambient noise. Moreover, it was shown that fixed and adaptive beamforming techniques are more suited for directional noise cases, where one interferer isolated in space degrades the clean speech signal. One approach could be to design a system that incorporates both coherence based filtering and beamforming methods to dynamically adapt to different noise scenarios; popular source localisation technique as well as coherence measures could be used to characterize the noise type and locate directional noise sources if such is the case. Adaptive noise cancelling could be run in parallel and given more importance when the interferers are indeed isolated and the adaptive filter converges. Such a system would probably perform better using a higher count of microphones.

7.2.3 Simplification in computational complexity
While the proposed system is not excessively demanding in terms of computational complexity, calculation of the EMSR suppression rules and the coherence post-filter could be simplified to ease integration in existing portable devices. At first, alternative approximations to the EMSR suppression rules that do not require calculation of the Bessel function [79] could be evaluated and their subjective performance studied, to see if they have a significant impact on speech quality. Calculation of the coherence over specific subbands, instead of the whole spectrum, could be implemented using successive QMF filtering and downsampling, thereby reducing complexity. The particular choice of subbands might also have an effect on the performance of the post-filter itself, as Bark spaced subbands were shown for example [20] to improve recovered speech quality in some cases.
8.1 Computation of window parameters

The windowing parameters used in eq. (2.2) can be computed as a function of the desired window length in seconds \( \text{window\_time} \), and the desired window overlap percentage \( 0 < \text{window\_overlap} < 1 \):

\[
N = \left\lfloor f_s \cdot \text{window\_time} \right\rfloor \tag{7.1}
\]

\[
S = \left\lfloor N \cdot (1 - \text{window\_overlap}) \right\rfloor \tag{7.2}
\]

This ensures integer valued \( N \) and \( S \) in samples, which is obviously required for any practical implementation. Since all processing of noisy speech file in this research project was done on a per-window basis, it was necessary to extract the right number of window out of the input to avoid any incomplete last window type of error. Thus, the number of valid, full windows that can be extracted from an input of length \( \text{input\_length} \) samples is defined as:

\[
\text{window\_count} = \left\lfloor \frac{\text{input\_length} - N}{S} \right\rfloor + 1 \tag{7.3}
\]

which means that at most \( N - 1 \) samples will not be processed from the input; this, again, is to ensure that all extracted windows are of the same size. If \( \text{window\_count} \) windows are extracted from the input, then the output length \( \text{output\_length} \) in samples will be defined by:

\[
\text{output\_length} = (\text{window\_count} - 1) \cdot S + N \tag{7.4}
\]

These definitions are used in the implementation to initialize windowing structures and output arrays.

8.2 Common window functions

Here we define common window functions of length \( N \) used in the analysis/synthesis part of speech enhancement algorithms, as described in Section 2.2:

74
\[ \text{square}[n] = 1 \]
\[ \text{hamming}[n] = 0.53836 - 0.46164 \cos \left( \frac{2\pi n}{N-1} \right) \]
\[ \text{hanning}[n] = 0.5 \left( 1 - \cos \left( \frac{2\pi n}{N-1} \right) \right) \] (7.5)
\[ \text{bartlett}[n] = \frac{2}{N-1} \left( \frac{N-1}{2} - \left| n - \frac{N-1}{2} \right| \right) \]

### 8.3 Derivation of the Wiener filter

The Wiener filter \( h[n] \) of order \( N \) seeks to minimize the MSE between an observed signal \( x[n] \) and a desired signal \( d[n] \). Formally, \( h[k] \) is solved by first posing the linear equation model:

\[ \hat{d}[n] = \sum_{k=0}^{N-1} h[k] x[n-k] \] (7.6)

defining the prediction error \( e[n] \) as:

\[ e[n] = \hat{d}[n] - d[n] \] (7.7)

and then making the error orthogonal to the observation (i.e. making cross-correlation between them null):

\[ E\{e[n]x[n-i]\} = 0 \quad i = 0 \ldots N-1 \] (7.8)

by simple substitution, we obtain:

\[ E\left\{ (\hat{d}[n] - d[n]) x[n-i] \right\} = 0 \]
\[ E\left\{ \left( \sum_{k=0}^{N-1} h[k] x[n-k] - d[n] \right) x[n-i] \right\} = 0 \] (7.9)
\[ \sum_{k=0}^{N-1} h[k] R_{xx}[i-k] = R_{xd}[i], \quad i = 0 \ldots N-1 \]

which is the general rule for solving the Wiener filter coefficients. This can be visualized in matrix form as:
\[
\begin{bmatrix}
R_\alpha[0] & \ldots & R_\alpha[-N+1] \\
\vdots & \ddots & \vdots \\
R_\alpha[N-1] & \ldots & R_\alpha[0]
\end{bmatrix}
\begin{bmatrix}
h[0] \\
h[N-1]
\end{bmatrix}
= 
\begin{bmatrix}
R_\alpha[0] \\
R_\alpha[N-1]
\end{bmatrix}
\] (7.10)

and thus \( h[k] \) can be solved by matrix inversion.

8.4 IIR Filter Block

It is often required to perform smoothing of either spectral amplitude or power. This is most easily realized with a simple 1st order IIR filter, such as:

\[
y[n] = ay[n-1] + (1-a)x[n]
\] (7.11)

where \( y[n] \) is a smoothed estimate of the observation \( x[n] \). Experimental values of \( \alpha \) yielding appropriate results for a specific application are often provided, but the value itself bares little information about the temporal characteristics of the estimate. When smoothing spectral coefficients, this depends on the sampling frequency \( f_s \) and the window step size \( S \).

To provide reproducible results that are independent of the window sizes, overlap and sampling frequency, we derive a formulation of \( \alpha \) that is a function of a time constant \( t_c \) only.

Taking the Z-transform of (7.11), and taking the unit step response \( x[n] \rightarrow u[n] \), we find that:

\[
y[n] = 1 - a^{n+1} = 1 - a^N
\] (7.12)

where \( N = n + 1 \) is the number of new observations \( x[n] \) at time \( n \). Substituting \( y[n] = 1 - e^{-1} \approx 0.63 \) (one time constant) in (7.12), we solve for \( a \) and obtain:

\[
a = e^{-1/N}
\] (7.13)

where \( N \), we recall, is the number of values of \( x[n] \) observed. Since we are normally dealing with spectral bins of consecutive spectral windows, we would like to substitute \( N \) for the number of spectral bins observed in a period of time \( t_c \). Since there are \( f_s/S \) windows per second:
such that \( a(t_c) \), the smoothing coefficient of spectral bins required for a one time constant response time of \( t_c \) seconds, is defined by:

\[
a(t_c) = e^{-S/(t_c f_s)}
\]  \hspace{1cm} (7.15)

This is illustrated by the filter block that implements (7.11) with \( a \rightarrow a(t_c) \):

\[t_c\]
\[\overrightarrow{x[n]}\]
\[\overrightarrow{IIR}\]
\[\overrightarrow{y[n]}\]

**Figure 8.1 IIR Filter Block**

It is assumed that \( y[n] \) actually feeds back into the block, as described by the recursive nature of the equation.

### 8.5 Derivation of the NLMS adaptive filter

The NLMS adaptive filter is illustrated in Figure 8.2.

\[\overrightarrow{x[n]}\]
\[\overrightarrow{y[n]}\]
\[\overrightarrow{h[k]}\]

**Figure 8.2 NLMS Adaptive Filter**

The system has inputs \( x[n] \) and \( y[n] \), and the NLMS algorithm seeks to minimize the expected energy of \( e[n] \) by adapting the coefficients of \( h[k] \). Segmenting the time signals as column vectors into windows of \( L \) samples, and writing for the \( k^{th} \) window in matrix notation:

\[
e_k[n] = x_k[n] - h_k^T y_k
\]  \hspace{1cm} (7.16)

\[
E_{NLMS,k} = E\left[ \left| e_k[n] \right|^2 \right] \approx e_k[n] e_k^*[n]
\]  \hspace{1cm} (7.17)
The approximation in (7.17) where the expectation operator is dropped is the simplification
done by the general class of LMS algorithms. The weights are then modified following the
rule of steepest descent:

$$h_{k+1} = h_k - \rho \frac{\partial E_{NMS,k}}{\partial h_k}$$  \hspace{1cm} (7.18)

Using (7.16) and (7.17) in (7.18), we can write:

$$h_{k+1} = h_k - \rho \frac{\partial}{\partial h_k} \left[ e_k[n] e_k^*[n] \right]$$
$$= h_k - \rho \frac{\partial}{\partial h_k} \left[ (x_k - h_k^T y_k)^T e_k[n] \right]$$  \hspace{1cm} (7.19)
$$= h_k - \rho y_k^T e_k[n]$$

which is the common LMS weight adaptation rule [78]. Since the weight difference between
two successive blocks in (7.19) (i.e. $|h_{k+1} - h_k|^2$) is proportional to the input signal energy
$|y_k|^2$, the step factor $\rho$ is set as to normalize the weight change between each step. Hence,
(7.19) is rewritten as:

$$h_{k+1} = h_k - \mu \frac{y_k^T e_k[n]}{y_k^T y_k + \varepsilon}$$  \hspace{1cm} (7.20)

where the small factor $\varepsilon << 1$ is introduced to avoid division by zero.

### 8.6 Error PSD and coherence measure

The link between the coherence of $x[n]$ and $y[n]$ and the PSD of their prediction error of the
ANC that adapts $y[n]$ to $x[n]$ is hereby developed. This error can be defined as:

$$e[n] = h[k]^* y[n] - x[n]$$  \hspace{1cm} (7.21)

which translates in the frequency domain to:

$$E(f) = H(f) Y(f) - X(f)$$  \hspace{1cm} (7.22)

Omitting frequency indexes for notation clarity, $P_{ee}(f)$ can be written as:
\begin{equation}
\begin{aligned}
P_{ee} &= E \left\{ \text{Re} \left( E \right) \right\} = E \left\{ \left[ H Y - X \right] \left[ H Y - X^* \right] \right\} \\
&= \left| H \right|^2 P_{yy} - H^* P_{yx} - H P_{yx}^* + P_{xx}
\end{aligned}
\end{equation}

(7.23)

Completing the squares in (7.23) and using the definition of the coherence in (3.24):

\begin{equation}
P_{ee} = \left| H - \frac{P_{yx}}{P_{yy}} \right|^2 P_{yy} - \frac{P_{yx}^2}{P_{yy}} + P_{xx}
\end{equation}

(7.24)

8.7 Distance measures

Here, we define several distance measures that are used in the objective quality evaluation of the implemented speech enhancement algorithms.

8.7.1 SNR measure

The SNR between $s[n]$ and $n[n]$ is normally defined as:

\begin{equation}
SNR_{\text{dB}} (s,n) = 10 \log_{10} \left( \frac{\sum_{n=1}^{N} s[n]^2}{\sum_{n=1}^{N} n[n]^2} \right)
\end{equation}

(7.25)

When evaluating speech enhancement algorithms, the noise $n[n]$ is understood to be the difference between the original speech and its time-aligned estimate:

\begin{equation}
SNR_{\text{dB}} (s, \hat{s}) = 10 \log_{10} \left( \frac{\sum_{n=1}^{N} s[n]^2}{\sum_{n=1}^{N} (s[n] - \hat{s}[n] - \Delta)^2} \right)
\end{equation}

(7.26)

The delay $\Delta$ is to chosen as to maximize $R_{sl}[n]$, such as:

\begin{equation}
\Delta = \arg \max_n R_{sl}[n]
\end{equation}

(7.27)
8.7 Distance Measures

8.7.2 Segmental SNR Measure

The SNR measure is known to be biased whereas strong signal sections affect the measure more than weak ones. As such, the segmental SNR measure was defined as:

$$SEGSNR_{\text{db}}(s, \hat{s}) = \frac{1}{M} \sum_{k=1}^{M} SNR_{\text{db}}(s_k, \hat{s}_k)$$ (7.28)

where the SNR measures are effectively averaged over the windowed segments of $s[n]$ and $\hat{s}[n]$. Note that the whole signal must again be time aligned before computing the $SNR_{\text{db}}$ measures.

8.7.3 Segmental SNR Improvement Measure

To quantify the improvement in segmental SNR of a speech enhancement algorithm, we take the ratio between the $SEGSNR_{\text{db}}$ at the output and at the input:

$$SEGSNR_{\text{IMPR,db}}(s, \hat{s}, n) = SEGSNR_{\text{db}}(s, \hat{s}) - SEGSNR_{\text{db}}(s, n)$$

$$= \frac{1}{M} \sum_{k=1}^{M} 10 \log_{10} \left( \frac{\sum_{n=1}^{N} n_k[n]^2}{\sum_{n=1}^{N} (s_k[n] - \hat{s}_k[n - \Delta])^2} \right)$$ (7.29)

$SEGSNR_{\text{IMPR,db}}$ is generally well correlated to the amount of perceived noise suppression of a speech enhancement algorithm.

8.7.4 LP-LSD measure

Not only do we need to measure the amount of perceived noise suppression, but it is also convenient to quantify the perceived speech distortion incurred by the algorithm. A popular method is to measure the log spectral distance between the LPC spectrum of the speech and speech estimate samples on a window basis, as the LPC spectrum is known to be well correlated to the 'perceived' speech spectrum.

In the Z-domain, we define the Z-transform of the $N^{th}$ order LPC approximation of $x[n]$ as:
\[ \mathbb{E}(\tilde{x}[n]) = \tilde{X}[z] = \frac{1}{1 + \sum_{k=1}^{N} a_{x,k} z^{-k}} \]  

(7.30)

where the \( a_{x,k} \) coefficients are found by solving the Wiener filter (Section 8.3) with \( x[n] \rightarrow x[n-k], \ k = 1 \ldots N \) and \( d[n] \rightarrow x[n] \). The prediction error of \( \tilde{x}[n] \) is defined as \( E_{\tilde{x}} \).

The discrete LPC spectrum \( \tilde{X}[m] \) is thus the DFT of \( \tilde{x}[n] \), and the LP-LSD can then be expressed as follows:

\[
LPLSD_{db}(s, \hat{s}) = \frac{1}{M} \sum_{k=1}^{M} \frac{1}{N} \sum_{l=1}^{N} \left| 20 \log_{10} \frac{\sqrt{E_{x_k}} \tilde{S}_x[l]}{\sqrt{E_{x_k}} \tilde{S}_x[l]} \right| \]

(7.31)

Naturally, a high LPLSD measure indicates a high level of speech distortion and vice versa. As an implementation detail, an LPC order of \( N = 24/16000 \cdot f_s \) is found convenient for proper spectrum approximation.
BIBLIOGRAPHY


9 BIBLIOGRAPHY


