Signal Subspace Speech Enhancement With Perceptual Post-Filtering

Mark Klein

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

May 2002

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering.

Abstract

Speech enhancement blocks form a critical part of voice communications systems. Unfortunately, most enhancement schemes have difficulty eliminating noise from speech without introducing distortion or artefacts. Many of the disturbances originate from poor parameter estimation and interframe fluctuations.

This thesis introduces the Enhanced Signal Subspace (ESS) system to mitigate the above problems. Based on a signal subspace framework, ESS has been designed to attenuate disturbances while minimizing audible distortion.

Artefacts are reduced by employing an auditory post-filter to smooth the enhanced speech spectra. This filter performs averaging in a manner that exploits the properties of the human auditory system. As such, distortion of the underlying speech signal is reduced.

Testing shows that listeners prefer the proposed algorithm to traditional signal subspace speech enhancement.

Sommaire

Les algorithmes qui améliorent la qualité de la voix constituent une partie essentielle des systèmes de communications modernes. Malheureusement, la plupart de ces techniques ont de la difficulté à éliminer le bruit sans introduire de distorsion ou d'artefacts. Plusieurs des ces perturbations proviennent de l'estimation imprécise des paramètres et des fluctuations intertrame du spectre. Ce mémoire présente le Système Enrichi de Sous Espace de Signal (ESS) pour limiter les problèmes énumérés ci-dessus. Basé sur les méthodes de sous-espace de signal, ESS a été conçu pour atténuer les perturbations tout en minimisant les déformations audibles. Les artefacts sont réduits en utilisant un post filtre auditif qui lisse le spectre de la parole. Pour filtrer sans introduire de distorsion, les propriétés du système auditif humain sont exploitées.

Les expériences montrent que les sujets préfèrent l'algorithme proposé aux méthodes de sous-espace de signal traditionnelles.

Acknowledgments

Firstly, I would like to thank my supervisor, Prof. Peter Kabal, for his invaluable advice and guidance. I have no doubts that my degree could not have been completed in a timely manner without his help and encouragement.

I further would like to express my gratitude to Tarun and Aziz for their help in proof-reading this thesis. My thanks are also extended to Chris and Hossein for their aid in understanding auditory masking theory. I wish to also thank Benoît for his help with the translation of my abstract and Alex for reading several sections of my results.

I am grateful to Wesley, Joachim and for their advice regarding speech enhancement. Their depth of knowledge in signal processing made them excellent resources. I would also like to acknowledge Paxton for his advice about structuring a thesis. And, I wish to thank Charif and Naveen for their support.

The AB tests could have not been carried out without the generous help of the volunteers who participated.

Finally, I would like to thank my parents for their love and support.

Contents

List of Figures

List of Tables

Definitions

Isomorphism A linear transformation $T : V \to W$ is called an *isomorphism* if it is both one-to-one and onto.

Metric Space A space X with metric d satisfying $\forall\, x, y, z \in X$

$$\begin{aligned} & d(x, y) \geq 0 \\ & d(x, y) = 0 \text{ iff } x = y \\ & d(x, y) = d(y, x) \\ & d(x, z) \leq d(x, y) + d(y, z) \end{aligned} \tag{1}$$

Null Space The *null space* of $\boldsymbol{A}$ is defined by [1]

$$\text{null}(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\} \tag{2}$$

Orthogonal Projector Let $S \subseteq \mathbb{R}^M$ be a subspace. $P \in \mathbb{R}^{M \times M}$ is the *orthogonal projection* onto S if $\text{ran}(P) = S$, $P^2 = P$, and $P^T = P$ [1].

Positive Definite Matrix A matrix $\boldsymbol{A}$ is *positive definite* if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{R}^n$ [1].

Range The *range* of $\boldsymbol{A}$ is defined by [1]

$$\text{ran}(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \text{ for some } \boldsymbol{x} \in \mathbb{R}^m\} \tag{3}$$

Subspace A subset W of a vector space V is called a *subspace* of V if W is itself a vector space under the addition and scalar multiplication defined in V [2].

Toeplitz Matrix A *Toeplitz matrix* is an $n \times n$ matrix $T_n = t_{kj}$ where $t_{kj} = t_{k-j}$.

Chapter 1

Introduction

The performance of voice communications systems degrades rapidly in adverse acoustic environments. For example, a user in a noisy automobile will be difficult to understand. Noise sources such as the engine, the wind, the ventilation system and the road [3] interfere with speech resulting in degraded speech quality and an overall loss of intelligibility. Speech enhancement systems endeavour to improve the performance of voice communication systems [4].

Enhancement will entail improving the quality and/or intelligibility of corrupted speech. To effect these changes, the technique of signal subspace speech enhancement will be employed. Furthermore, a perceptual post-filter will be employed to remove artefacts introduced by enhancement.

This chapter will introduce the concept of signal subspace speech enhancement. This will be followed by a brief review of the history of signal subspace methods. Finally, the motivation for an enhanced signal subspace speech enhancement algorithm will be presented.

1.1 Applications of Speech Enhancement Algorithms

Many voice communication systems require speech restoration blocks to function properly. Telecommunication systems form the primary application of speech enhancement systems, but many others exist. These include hearing aids and damaged recordings.

Telecommunication systems do not perform well in noisy environments. Ambient noise prevents the speech coding blocks from accurately estimating the required spectral parameters. Thus, the resulting coded speech sounds mechanical and distorted. In addition, it still contains the corrupting noise. To improve performance, a speech enhancement system can be placed as a front end to reduce noise energy [5–8].

Speech enhancement is vital in hearing aids. These devices help the hearing impaired by amplifying ambient audio signals. Unfortunately, noise content is increased along with the speech, reducing the intelligibility of the signal presented to the user. To improve speech quality, a speech enhancement block may be utilized as a pre-processing stage [9, 10].

It is the goal of audio restoration to remove any audio object, which is not part of the intended recording. This includes disturbances introduced by the storage media and environmental noise recorded. O'Shaughnessy *et al.* utilized speech enhancement techniques to improve the intelligibility of a wire-tap recording in [11].

1.2 Signal Subspace Approach for Signal Enhancement

The removal of additive noise from speech has been an active area of research for several decades. Numerous methods have been proposed by the signal processing community. Among the most successful signal enhancement algorithms have been spectral subtraction [12, 13] and Wiener filtering [14]. However, these algorithms tend to introduce artefacts (disturbances) and distortion. Signal subspace enhancement techniques have been shown to insert fewer disturbances.

Signal subspace speech enhancement techniques decompose the input into signal components and noise components. To improve speech quality, the noise components are discarded. If the decomposition is performed correctly, the amount of noise in the speech signal should decrease without creating distortion. The improved speech signal can be further processed for better quality.

Many signals of interest span a reduced dimensionality subspace of a larger complex vector space. When noise is added to a signal, the resultant vector will be perturbed outside the subspace. Removing content that does not lie within the reduced dimensionality subspace will improve signal quality. Signal subspace enhancement techniques attempt to decompose a vector space into two subspaces: a signal subspace and a noise subspace [15]. Enhancement can then be performed by discarding the noise subspace and estimating the clean signal from the remaining content within the signal subspace [16]. The decomposition has typically been done using the Karhunen-Loève (KL) expansion or Singular Value

Decomposition (SVD).

In the past, signal subspace processing had been applied to the direction of arrival problem and the detection of sinusoids in noise. However, Dendrinos *et al.* [17] applied this methodology to speech enhancement. As speech can be well represented with a simple linear model, it was an excellent candidate for signal subspace enhancement.

1.3 History of Signal Subspace Speech Enhancement Techniques

Signal subspace techniques have matured significantly over the last thirty years. They originated with Pisarenko's work involving the detection of sinusoids in noise [18]. The method was adapted to process a more general class of signals. Further work attempted to reduce computational complexity, provide handling of coloured noise and incorporate perceptual modelling.

The development of signal subspace techniques will focus on two areas: detection of sinusoids and enhancement of speech. The following will place an emphasis on the KL expansion implementation.

1.3.1 Signal Subspace Techniques for Sinusoidal Signals

In 1973, Pisarenko developed a method of detecting p sinusoids in additive white noise [18]. This algorithm required knowledge of the sinusoid frequencies and the $(2p+1) \times (2p+1)$ covariance matrix of the noisy signal.

Multiple Signal Classification (MUSIC) was developed by Schmidt [19] to determine the parameters of multiple wavefronts arriving at an antenna array. MUSIC was an improvement over Pisarenko's method as it could detect the frequencies of the transmitted sinusoids.

Tufts *et al.* presented a method for retrieving the signal component from a noisy data set [20]. Their method entailed creating a Hankel data matrix, calculating the SVD and nulling the singular values corresponding to the noise signal alone. This approximation method was known as least squares estimator (LS) as it returned the projection of the noisy signal onto the signal subspace.

De Moor introduced the minimum variance estimator in [21]. This estimator sought to minimize the mean square error of the reconstructed signal. In this paper, De Moor also showed that it was impossible to recover the original noise column space of the exact data signal using the method Tufts *et al.* presented in [20]. He proved that the angle between the true and estimated subspaces would always be a function of the signal-to-noise ratio.

1.3.2 Signal Subspace Techniques for Speech Signals

Dendrinos *et al.* first utilized signal subspace techniques to enhance speech in [17]. Furthermore, they introduce a method of estimating the dimensionality of the signal subspace.

Ephraim and Van Trees used the KL expansion for signal decomposition [16]. They also proposed two signal estimators: the spectral domain constraint (SDC) and the time domain constraint (TDC). The former attempted to spectrally shape the residual noise while the latter constrained residual noise energy.

Huang and Zhao proposed further enhancements to the KL expansion method proposed by Ephraim and Van Trees. In [22], they discussed an energy-constrained signal subspace method (ECSS). The key concept was to match the short-time energy of the enhanced speech signal to the unbiased estimate of the clean speech. They asserted that this method was effective in recovering the low-energy segments in continuous speech. In addition, Huang and Zhao showed that a discrete cosine transform could be used as a substitute to the KL expansion in their ECSS algorithm [23]. This reduced computational complexity from $O(N^3)$ to $O(N^2)$.

Rezayee and Gazor [24] incorporated coloured noise handling into their algorithm by diagonalizing the noise correlation matrix using the estimated eigenvalues of the clean speech and nulling any off-diagonal elements. In addition, they incorporated subspace using the projection approximation algorithm developed by Yang [25].

In [26], Mittal and Phamdo proposed a new approach for enhancing speech degraded by coloured noise. Noisy speech frames were classified as speech-dominated frames and noise-dominated frames. In the speech dominated frames, the estimated signal correlation matrix is used to calculate the KL expansion, otherwise, the noise correlation matrix is employed.

Recently, Jabloun introduced a method to incorporate the masking properties of the ear [27]. In this publication, the Wiener filter coefficients were calculated using eigenvalues that correspond to the noisy excitation pattern. These eigenvalues were determined by projecting the excitation pattern of the noisy signal onto the squared magnitude of the individual eigenvectors.

1.4 Description of Thesis Work

This thesis will study the benefits of utilizing a perceptual post-filter to smooth the output of signal subspace speech enhancement systems. By using knowledge of the human ear, it is expected that artefact suppression may be accomplished without significantly distorting the underlying speech signal. This enhancement method will be denoted as the Enhanced Signal Subspace (ESS) method. A block diagram of the proposed system is shown below.

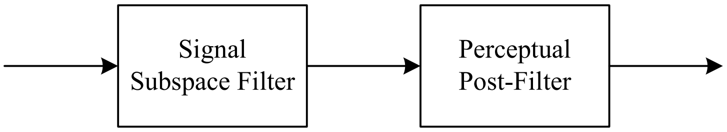


Fig. 1.1 Overview of Enhanced Signal Subspace speech enhancement system

Most enhancement methods rely on estimates of second-order statistics of the noise and speech signals to calculate filter gains. Due to the errors inherent in the measurement process, audible artefacts known as *musical noise* are invariably introduced into the enhanced speech. Musical noise is an auditory disturbance resembling a sum of sinusoids of changing frequencies, turning off and on from frame-to-frame. It is the most common artefact associated with speech enhancement.

Signal subspace methods improve estimates of signal parameters by averaging over long windows. However, this does not result in complete elimination of musical noise. While artefacts will no longer originate from fluctuations in the noisy spectrum estimator, new sources emerge. These include rapid changes of model order and subspace swapping. The latter condition refers to noise basis vectors being incorrectly employed to describe the signal subspace.

A great deal of effort has been expended on the development of techniques to eliminate musical noise. Most schemes utilize forms of temporal and spectral averaging to lessen its presence. Several new suppression methods incorporate knowledge of perception in the design of the enhancement filter. It has been shown in numerous papers [28, 29] that perceptual filters are a potent tool for eliminating musical noise. By employing a filter based on the notion of auditory masking, musical noise will be reduced with minimal distortion.

1.5 Organization of Thesis

This thesis is divided into six chapters, including this introduction.

Chapter 2 presents the fundamentals of signal subspace speech enhancement. Starting with a brief review of the properties of speech, it proceeds to describe the signal subspace speech enhancement algorithm. Attention is also placed on subspace dimensionality estimation and coloured noise compensation.

Chapter 3 describes the operation of the perceptual post-filter. A description of the phenomenon of musical noise is presented. The concept of auditory masking and the PEAQ masking model is then introduced.

Chapter 4 details the design of the perceptual post-filter. Motivation for the utilization of auditory filters for musical noise suppression is provided. Afterwards, an overview of the ESS system is presented. Finally, a detailed description of the functional blocks is provided.

Chapter 5 contains a performance analysis of the proposed algorithm. An examination of the implementation issues is presented. A qualitative examination of the properties of the ESS enhancement method follows. Then, the results of objective and subjective testing are discussed.

Lastly, Chapter 6 summarizes this thesis and presents directions for future work.

Chapter 2

Signal Subspace Based Speech Enhancement

Signal subspace based speech enhancement techniques decompose M-dimensional spaces into two subspaces: a signal subspace and a noise subspace. It is assumed that the speech signal can lie only within the signal subspace while the noise spans the entire space. Only the contents of the signal subspace are used to estimate the original speech signal.

This chapter will describe the process of decomposing the complex space into orthogonal subspaces and describe several estimators which have been applied in previous work.

2.1 Problem Description

The speech enhancement problem will be described as a speech signal $\boldsymbol{x}$ being transmitted through a distortionless channel that is corrupted by additive noise $\boldsymbol{w}$. The resulting noisy speech signal $\boldsymbol{y}$ can be expressed as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w} \tag{2.1}$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]^T$, $\boldsymbol{w} = [w_1, w_2, \ldots, w_M]^T$ and $\boldsymbol{y} = [y_1, y_2, \ldots, y_M]^T$. The observation period has been denoted as M. Henceforth, the vectors $\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}$ will be considered as part of $\mathbb{C}^M$.

The speech enhancement system will attempt to estimate the original signal using a single channel of received speech.

2.1.1 Speech and Noise Requirements

The following assumptions are made about the speech and noise signals. It should be noted that these requirements are sufficiently weak that a large class of signals can be accommodated by the signal subspace algorithm.

- Noise and speech are zero mean random processes
- Frames of speech are incrementally Wide-Sense Stationary (WSS): This supposition is based on the physiology of the human speech organs. The vocal tract and excitation vary slowly over time. Over the course of a long vowel, a window as large as 100 ms can be used without obscuring the desired patterns via averaging [30]. In this work, it will be assumed that a speech frame of up to 50 ms will be wide-sense stationary.
- Noise and speech are orthogonal: It will be assumed that the noise signal is uncorrelated with the speech signal. Thus, $E\{\boldsymbol{x}\boldsymbol{w}^H\} = \mathbf{0}$. As the noise and speech sources are zero mean and independent random processes, this condition is satisfied.
- Noise is a white random process: The noise will be modelled as an uncorrelated random process with variance σ_w^2. Therefore,

  $$\boldsymbol{R}_w = E\{\boldsymbol{w}\boldsymbol{w}^H\} = \sigma_w^2 \boldsymbol{I}. \tag{2.2}$$

  Coloured noise can be rendered white via the application of a prewhitening filter (see Section 2.8).
- All signals are correlation ergodic: It will be assumed that the signals that are under analysis will be correlation ergodic. As such, the time average converges to the expected value as the observation values become large. Thus,

  $$\lim_{N\to\infty} \frac{1}{N} \sum_{i=-N}^{N} \bar{x}_i x_{i+m} = r_{x_{mm}} \tag{2.3}$$

  where $r_{\bar{x}_{mm}} = E\{\bar{x}_n x_{n+m}\}$ and $\bar{\cdot}$ denotes the conjugation operation.

2.2 A Brief Background on Speech

This section will treat several aspects of speech production, as well as, articulatory and acoustic phonetics [30].

2.2.1 Speech Production

The speech organs can be divided into three main groups: the lungs, the larynx and the vocal tract. The vocal tract is comprised of the oral pharyngeal cavities. The speech organs are depicted in Fig. 2.1.

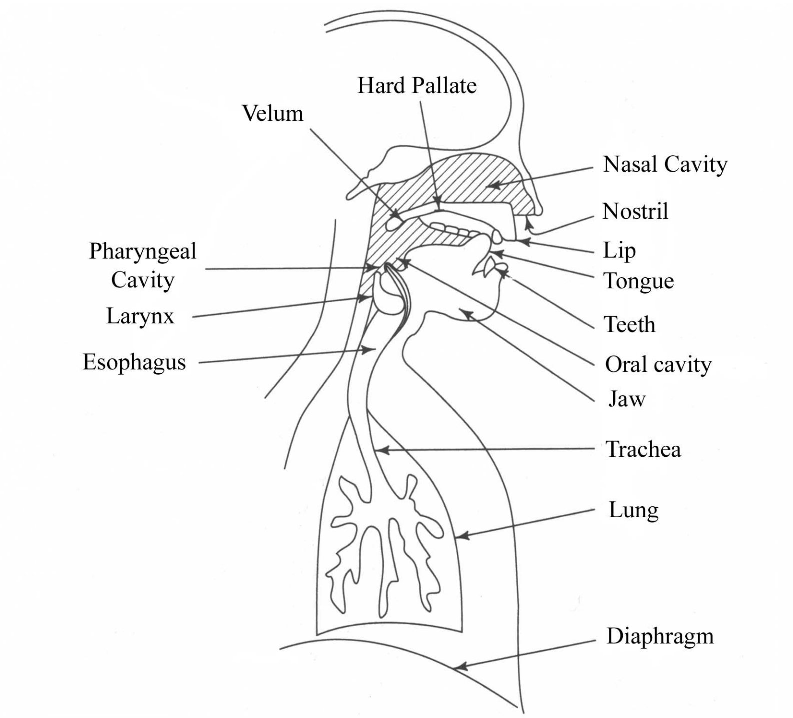


Fig. 2.1 Cross-sectional view of the speech organs, from [31]

The lungs provide the source of airflow which passes through the vocal tract. Normal breathing creates little audible sound because air expelled by the lungs passes unobstructed through the vocal tract. As pressure varies, sound occurs when the airflow path is narrowly constricted or totally occluded, interrupting the airflow to create either noise or pulses of air.

The role of the larynx is to produce a periodic excitation for the vocal tract. The larynx contains the vocal folds, a pair of elastic structures of tendon, muscles, and mucous membranes. They open and close at a rate known as the *fundamental frequency*.

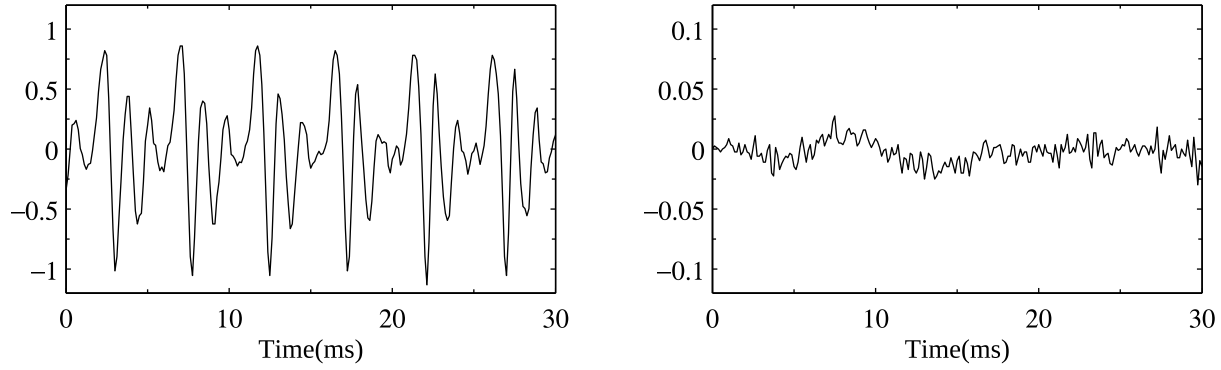
The vocal tract has two specific functions: It can modify the spectral distribution of energy in glottal waveforms and it can contribute to the generation of sound for obstruent sounds. Structures in the vocal tract that move are known as articulators. The most important articulators include the tongue, lips, velum and larynx. The vocal tract can be modelled as an acoustic tube. The resonant frequencies of the vocal tract are known as formants. Typically, a vowel will have only 3–5 formants within its bandwidth.

2.2.2 Articulatory/Acoustic Phonetics

Speech can be identified as either voiced or unvoiced. These two classes differ in articulation, duration and intensity.

A strong periodic waveform is attributable to voiced speech. In this case, the vocal tract is clear of obstructions and the glottis vibrates in a regular manner. This group includes vowels, dipthongs, glides and nasals. It can be noted that some fricatives may have a voicing component.

Unvoiced speech is characterized by an aperiodic waveform. Physically, it results from air passing through a stationary glottis to an obstruction. The resulting turbulence produces a noise-like output. Some fricatives and stops fall under this grouping.



(a) Voiced speech

(b) Unvoiced speech

Fig. 2.2 Comparison between voiced and unvoiced speech

Phonemes are typically classified in terms and manner and place of articulation. Manner of articulation refers to the manner in which the vocal tract is obstructed. Place of articulation refers to the place where the occlusion can occur. Places of articulation include the labials, the velum and the teeth.

Vowels are the most intense type of phonemes with durations varying from 50 to 400 ms. The frequency band beneath 1000 Hz contains the majority of the phonemes spectral energy. Vowels can be distinguished by the first three formants. There is a -6 dB/octave drop in energy with frequency. The vocal tract is unobstructed during vowels. Spectrum shaping is performed by the tongue and lips.

Fricatives tend to have aperiodic waveforms. Unlike vowels, the majority of fricatives' energy is concentrated in the higher frequency bands. Fricatives are characterized by a major obstruction in the vocal tract. In the unvoiced case, the noise source is located anterior to the major constriction. When fricatives are voiced, they are characterized with a low frequency formant (around 150 Hz) known as a voice bar. Additionally, some voiced fricatives will have weak harmonics at lower frequencies.

Stops are produced by completely obstructing the vocal tract at some point, allowing pressure to build up, then releasing the pressure suddenly. A noise burst first ensues, exciting all frequencies but primarily those which correspond to the fricative the vocal tract is matching. Stops are transient signals with an average duration of 10 ms.

2.2.3 Low-Rank Modelling of Speech

Low-rank modelling refers to the process where a data space is transformed into a feature space that, in theory, has the same dimension as the original data [32]. It has been established from previous work in speech compression and speech enhancement that such a representation exists and that the underlying speech signal is well represented.

The utilization of reduced dimensionality in speech compression has been successfully employed to increase coding gain. Two successful applications of this paradigm include sinusoidal modelling and wavelet compression. Sinusoidal modelling attempts to model the excitation of the vocal tract using a sum of sinusoids with different amplitudes, phase and frequency [33, 34]. Wavelet compression discards weaker coefficients resulting from a discrete wavelet transform [35, 36].

Though low rank representations are effective, speech does, in fact, have full rank. This statement can be verified by plotting the eigenvalues of the frame correlation matrix of a speech signal. Such a representation is known as a *Scree graph* [37]. Since the eigenvalue matrix resulting from the speech correlation matrix is diagonal, the number of nonzero eigenvalues indicate the rank.

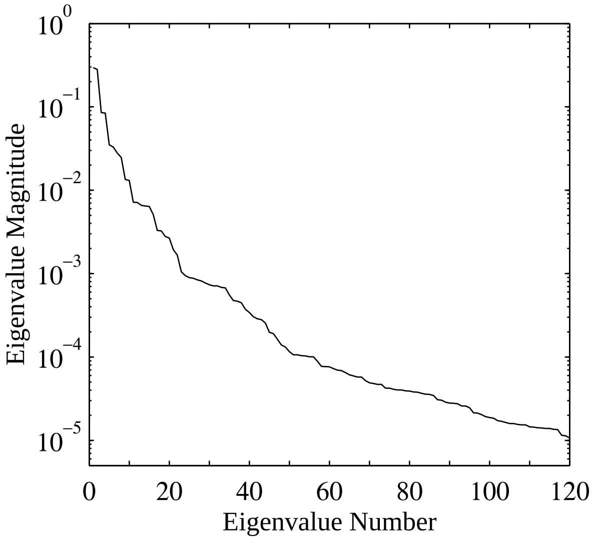


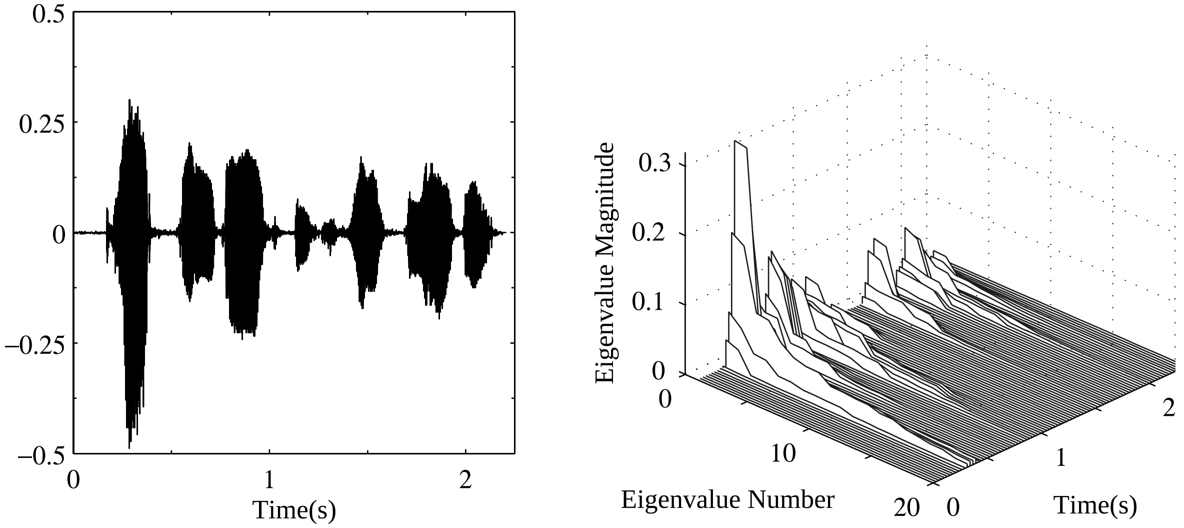
Fig. 2.3 Scree graph

Fig. 2.3 shows the plot of a typical frame of speech. A 300 sample rectangular data window was utilized to estimate the 120 lag correlation matrix. It should be noted that the eigenvalues have been sorted in descending order. As all eigenvalues are nonzero, the correlation matrix is not rank deficient.

Further information can be obtained by plotting the Scree graphs of successive frames. Fig. 2.4 is the concatenation of a series of scree plots for a speech signal. The utterance "Cats and dogs each hate the other" was used. Correlation matrices with 40 lag were employed. They were estimated using 300 sample rectangular windows of data with 50 % overlap between frames.

In voiced sections, the majority of the signal energy can be modelled using a few eigenvectors. This should not be surprising due to the highly correlated nature of periodic speech.

Conversely, the eigenvalues of the unvoiced section are more uniform. Thus, a higher order model will be required to model these utterances with the same degree of accuracy as voiced speech.



(a) Time signal

(b) Plot of signal spectra

Fig. 2.4 Time and eigendomain representations of a speech utterance

2.3 Concept of Signal and Noise Subspaces

If it is assumed that speech signals are confined to a subspace of dimensionality K, where $K < M$, then $\mathbb{C}^M$ can be decomposed into two subspaces: a *signal subspace* and a *noise subspace*. The *signal subspace* will correspond to the reduced dimensionality subspace where speech may exist. Meanwhile, the *noise subspace* will only contain noise.

Ephraim and Van Trees [16] realized this partitioning by postulating a linear model for the speech frame under analysis. The range and the null space were characterized as the signal and noise subspaces respectively.

2.3.1 General Linear Speech Model

The linear model for the clean signal assumes that every length M frame can be represented using the model

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{s} = \sum_{i=1}^{K} s_i \boldsymbol{v}_i \qquad K \leq M \tag{2.4}$$

where $\boldsymbol{s} = [s_1, s_2, \ldots, s_K]^T$ is a sequence of zero mean complex random variables. $\boldsymbol{V} \in \mathbb{R}^{M \times K}$ is known as the *model matrix*. Assuming that the columns of $\boldsymbol{V}$ are linearly independent, then $\boldsymbol{V}$ will have a rank of K. The range of $\boldsymbol{V}$ defines the signal subspace. It will henceforth be denoted as V.

While a linear model with rank M is certainly possible, it will be assumed that K is strictly less than M. Otherwise, V would contain all of $\mathbb{C}^M$ and no orthogonal noise subspace would exist.

The noise subspace will be denoted as $V^{\perp}$. With dimension $M-K$, it is the null space of the model matrix. This subspace only contains vectors resulting from the noise process. The union of V and $V^{\perp}$ span $\mathbb{C}^M$.

It should be reinforced that as the noise process has full rank, it spans $\mathbb{C}^M$.

2.4 Karhunen-Loève Expansion Based Linear Model

The Karhunen-Loève (KL) expansion has had many applications in communications, image compression and statistical analysis. It will be demonstrated that the KL expansion is the optimal[1] basis for signal subspace decomposition.

2.4.1 Fundamentals of the Karhunen-Loève Expansion

It has been shown in many applications that the KL expansion is an excellent basis for dimensionality reduction. The following definition is from Haykin [32]:

Definition 1 (Karhunen-Loève Expansion) *Let the M-by-1 vector $\boldsymbol{u}$ denote a data sequence drawn from a wide-sense stationary process of zero mean and correlation matrix $\boldsymbol{R}_u$. Let $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_M$ be eigenvectors associated with the M eigenvalues of the matrix $\boldsymbol{R}_u$. The vector $\boldsymbol{u}$ may be expanded as a linear combination of these eigenvectors as follows*

$$\boldsymbol{u} = \sum_{i=1}^{M} c_i \boldsymbol{q}_i. \tag{2.5}$$

The coefficients of the expansion are zero-mean, uncorrelated random variables defined by

[1] In the mean-square error sense.

the inner product

$$c_i = \boldsymbol{q}_i^H \boldsymbol{u}. \tag{2.6}$$

It can be shown that the KL expansion will always exist for a WSS random process using the spectral theorem.

Theorem 1 (Spectral Theorem) *Every Hermitian matrix can be diagonalized by a unitary matrix* $\boldsymbol{Q}$

$$\boldsymbol{A}=\boldsymbol{A}^H \;\Rightarrow\; \boldsymbol{A}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^H \tag{2.7}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_M)$.

Such a representation is known as a Schur decomposition.

Clearly, as all WSS processes have Hermitian correlation matrices, they are diagonalizable. Even, if the correlation matrix is singular, the KL expansion will still exist. However, the column vectors of $\boldsymbol{Q}$ will not be linearly independent.

2.4.2 Subspace Decomposition Using Karhunen-Loève Expansion

If an eigendecomposition is performed on the correlation matrix of the speech signal $\boldsymbol{x}$, the following form is obtained

$$\boldsymbol{R}_x = \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{x}_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_1^H \\ \boldsymbol{Q}_2^H \end{bmatrix} \tag{2.8}$$

where $\boldsymbol{\Lambda}_{x_1} = diag(\lambda_{x_1}, \ldots, \lambda_{x_K})$.

The eigenvector matrix $\boldsymbol{Q}$ has been partitioned into two sub-matrices, $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$. The matrix $\boldsymbol{Q}_1$ contains eigenvectors corresponding to non-zero eigenvalues. These eigenvectors form a basis for the signal subspace. Meanwhile, $\boldsymbol{Q}_2$ contains the eigenvectors which span the noise subspace.

The matrix $\boldsymbol{Q}_1\boldsymbol{Q}_1^H$ is idempotent ($\boldsymbol{P}^2 = \boldsymbol{P}$), Hermitian and $span(\boldsymbol{Q}_1) = span(\boldsymbol{V})$. Thus, $\boldsymbol{Q}_1\boldsymbol{Q}_1^H$ is a projector onto the signal subspace. Similarly, $\boldsymbol{Q}_2\boldsymbol{Q}_2^H$ is the projector onto the noise subspace. As both subspaces complete $\mathbb{C}^M$, any input vector can be represented as

$$\boldsymbol{u} = \boldsymbol{Q}_1\boldsymbol{Q}_1^H\boldsymbol{u} + \boldsymbol{Q}_2\boldsymbol{Q}_2^H\boldsymbol{u}. \tag{2.9}$$

The expected power of a Karhunen-Loẽve coefficient can be shown to be equal to

$$E\{c_i^2\} = \lambda_i. \tag{2.10}$$

As the eigenvectors which make up $\boldsymbol{Q}_2$ have null eigenvalues, they contribute no energy to the speech signal. As such, they can be omitted in a KL expansion without introducing error. The noise subspace eigenvectors, corresponding to a zero eigenvalue with multiplicity $M-K$, apart from being orthogonal to each other, are arbitrary [37].

Thus, a reduced rank representation for the signal $\boldsymbol{u}$ will have the form

$$\tilde{\boldsymbol{u}} = \sum_{i=1}^{K} c_i \boldsymbol{q}_i = \boldsymbol{Q}_1 \boldsymbol{c}. \tag{2.11}$$

2.4.3 Optimal Low-Rank Representation

The truncated expansion presented in Eq. (2.11) has the property of being the optimal low-rank representation for an arbitrary WSS random process. Stated succinctly, the KL expansion satisfies

$$\min_{\substack{\phi_1,\ldots,\phi_K \\ \boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_K}} E\{\|\boldsymbol{x} - \sum_{i=1}^{K} \phi_i(\boldsymbol{x})\boldsymbol{\nu_i}\|_2^2\} \tag{2.12}$$

where $K \leq M$, $\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_K$ are arbitrary vectors in $\mathbb{R}^M$ and $\phi_1, ..., \phi_K$ are arbitrary functionals $\mathbb{R}^M \rightarrow \mathbb{R}$. If the rank of a signal is underestimated, the truncated KL expansion minimizes the mean-square error (MSE) of the representation. For a proof of this property, see [38].

The energy of the error associated with a KL representation truncated to K length can be evaluated as

$$\epsilon_u = E\{\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|_2^2\} = \sum_{i=K+1}^{M} \lambda_i. \tag{2.13}$$

The KL expansion is often referred to as the projection matrix onto the rank-M principal subspace [39] or principal component analysis.

2.5 Subspace Estimation From Noisy Data

Estimating the signal subspace from the original speech file is straightforward. However, when noise has been added to the signal, additional considerations must be addressed. It will be shown how to estimate the signal speech correlation matrix using noisy data.

2.5.1 Estimation of Signal Correlation Matrix

The correlation matrix of the noisy speech signal can be expanded as

$$\begin{aligned}\boldsymbol{R}_y &= E\left\{\boldsymbol{y}\boldsymbol{y}^H\right\} \\ &= \boldsymbol{R}_x + \boldsymbol{R}_w \\ &= \boldsymbol{R}_x + \sigma_w^2 \boldsymbol{I}.\end{aligned} \tag{2.14}$$

Accordingly, the correlation matrix of the original speech signal can be calculated by

$$\boldsymbol{R}_x = \boldsymbol{R}_y - \sigma_w^2 \boldsymbol{I}. \tag{2.15}$$

The parameters $\boldsymbol{R}_y$ and σ_w^2 are typically estimated. The quality of the estimates directly affects the accuracy of the calculated eigenvalues and eigenvectors.

2.5.2 Sensitivity Analysis on Eigenvalue Problem

To determine the effect that measurement error (via imperfect estimators) and machine precision have on the subspace estimates, a sensitivity analysis will be performed using classical eigenvalue theory. It will be assumed that a matrix $\boldsymbol{A}$ has been perturbed by a matrix $\epsilon\boldsymbol{B}$ resulting in $\boldsymbol{C}$. Therefore,

$$\boldsymbol{C} = \boldsymbol{A} + \epsilon\boldsymbol{B} \tag{2.16}$$

Eigenvalue Problem

A bound on the error resulting from perturbing the correlation matrix can be obtained by examining the conditioning of the general eigenvalue problem [40]. Let $\hat{\lambda}_i$ denote the calculated value of the eigenvalue λ_i.

$$|\lambda_i - \hat{\lambda}_i| \leq \epsilon\kappa(\boldsymbol{Q}) \left\|\boldsymbol{B}\right\|_2 \tag{2.17}$$

where $\kappa(\boldsymbol{D}) = \|\boldsymbol{D}^{-1}\|_2 \|\boldsymbol{D}\|_2$ denotes the condition number. The quantity $\kappa(\boldsymbol{Q})$ can be expressed as

$$\begin{aligned} \kappa(\boldsymbol{Q}) &= \|\boldsymbol{Q}^{-1}\|_2 \|\boldsymbol{Q}\|_2 \\ &= \|\boldsymbol{Q}^{H}\|_2 \|\boldsymbol{Q}\|_2 \\ &= 1 \end{aligned} \tag{2.18}$$

Therefore, the eigenvalue problem will always be well-conditioned when considering correlation matrices produced by a WSS random process. Accordingly, the error can be bounded as

$$|\lambda_i - \hat{\lambda}_i| \leq \epsilon \, \|\boldsymbol{B}\|_2 \tag{2.19}$$

Eigenvector Problem

A first-order estimate of the perturbation bound can be estimated by Eq. (2.20) [40]

$$|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i| \leq \epsilon \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\beta_{ij} x_i}{(\lambda_i - \lambda_j)} \tag{2.20}$$

where $\beta_{ij} = \boldsymbol{x}_i^T \boldsymbol{B} \boldsymbol{x}_j$. This bound assumes that the eigenvectors under consideration correspond to distinct eigenvalues. This assumption will hold true for the eigenvectors corresponding to the signal subspace. Clearly, we can see that if two or more eigenvalues are close to each other, the corresponding eigenvectors are very sensitive to perturbations.

2.6 Rank Estimation

Estimating the order of a speech correlation matrix perturbed by additive noise is difficult. Three algorithms are presented that have been shown to be effective in adverse conditions. The dimensionality of the signal subspace should be chosen to discard a large amount of noise energy while preserving the quality of the speech signal.

An example of an order estimate with clean speech is provided in Fig. 2.5. The rank has been chosen manually using the given paradigm.

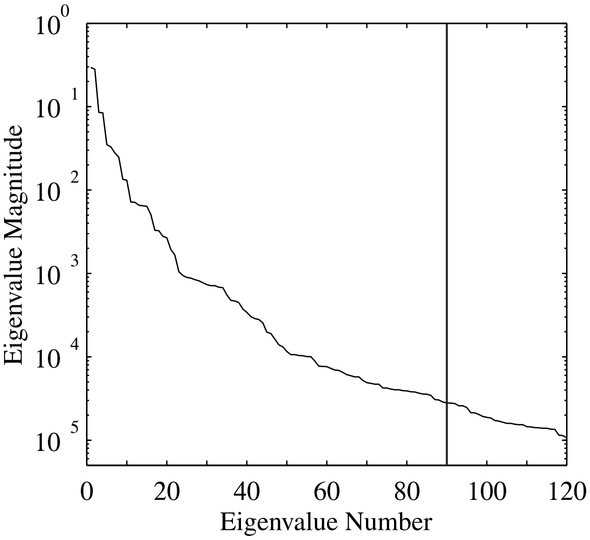


Fig. 2.5 Order estimate

2.6.1 Theoretical Estimator

The theoretical estimator assumes that the order of the system equals the number of noisy correlation matrix eigenvalues that exceed the variance of the noise. This ensures that poorly estimated eigenvalues and eigenvectors are not used to define the signal subspace. Hence,

$$K^* = \#\{k \in \mathbb{Z}_+ : \lambda_{y_i} > \sigma_w^2\}. \tag{2.21}$$

This method of estimating the rank is clearly computationally efficient though it is not optimal for full-rank signals. It also does not attempt to make a trade-off between noise removal and signal distortion.

2.6.2 Minimum Description Length

The minimum description length (MDL) utilizes a simple criterion to deduce the order of a system: Choose a model that minimizes the coded length of the observations. Rissanen justified this rationale in [41] by arguing that maximum compression of a sequence is achieved when the statistical properties of the data are utilized. Thus, the model order that best characterizes a system is the one that permits the shortest representation.

Rissanen defined the expected description length as [42]

$$L(i) = L(\boldsymbol{y}|\boldsymbol{\theta}(i)) + L(\boldsymbol{\theta}(i)). \tag{2.22}$$

The first term in the MDL-estimator acts as a measure of the expected codeword length of the parameterized signal. The second term can be interpreted as the penalty associated with communicating the model [43].

The KL expansion model is characterized by the i^{th} order parameter vector $\boldsymbol{\theta}(i)$

$$\boldsymbol{\theta}(i) = \begin{bmatrix} \lambda_{y_1}^T & \ldots & \lambda_{y_i}^T & \sigma_w^2 & \boldsymbol{q}_1^T & \ldots & \boldsymbol{q}_i^T \end{bmatrix}^T \tag{2.23}$$

The expected codeword length for the parameterized signal can be calculated using the negative log likelihood. It will be assumed that B observations are available to determine model order.

$$L(\boldsymbol{y}|\boldsymbol{\theta}(i)) = -\log\{f\{\boldsymbol{y}(1), \ldots, \boldsymbol{y}(B)|\boldsymbol{\theta}(i)\}\} \tag{2.24}$$

The speech signal $\boldsymbol{x}$ will be modelled by a zero mean Gaussian distribution. Accordingly, $\boldsymbol{y}$ will be a zero mean Gaussian random variable. Using this premise, $L(\boldsymbol{y}|\boldsymbol{\theta}(i))$ can be shown to be equal to

$$L(\boldsymbol{y}|\boldsymbol{\theta}(i)) = -B\log\{\det\{\boldsymbol{R}_y(i)\} - \operatorname{tr}\{\{\boldsymbol{R}_y(i)\}^{-1}\hat{\boldsymbol{R}}_y\} \tag{2.25}$$

where $\hat{\boldsymbol{R}}_y$ is the sample covariance matrix

$$\hat{\boldsymbol{R}}_y = \frac{1}{B}\sum_{j=1}^{N} \boldsymbol{y}(j)\boldsymbol{y}^H(j). \tag{2.26}$$

Further it will be assumed that $\boldsymbol{R}_y(i)$ can be expanded as:

$$\boldsymbol{R}_y(i) = \boldsymbol{Q}\begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{x}_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{Q}^H + \sigma_w^2\boldsymbol{I} \tag{2.27}$$

where $\boldsymbol{\Lambda}_{x_1} = diag(\lambda_{x_1}, \ldots, \lambda_{x_i})$.

The i^{th} order parameter vector must be estimated from the B observations. Thus, approximations of the eigenvalues, eigenvectors and noise variance are obtained using the maximum likelihood (ML) estimators [44]

$$\hat{\lambda}_{y_j} = l_j \qquad j = 1, \ldots, i \tag{2.28a}$$

$$\hat{\sigma}_w^2 = \frac{1}{M-i} \sum_{j=i+1}^{M} l_j \tag{2.28b}$$

$$\hat{\boldsymbol{q}}_i = \boldsymbol{c}_i \tag{2.28c}$$

where $l_1 > \ldots > l_M$ and $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M$ are the eigenvalues and eigenvectors of the sample covariance matrix $\hat{\boldsymbol{R}}_y$. Substituting the maximum likelihood estimates into Eq. (2.25), the following is obtained

$$L(\boldsymbol{y}|\boldsymbol{\theta}_{ML}(i)) = -(M-i)B\Big(\log\Big\{\prod_{j=i+1}^{M} \lambda_{y_j}^{\frac{1}{M-i}}\Big\} - \log\Big\{\frac{1}{M-i}\sum_{j=i+1}^{M} \lambda_{y_j}\Big\}\Big). \tag{2.29}$$

Returning to the second term of Eq. (2.22), the description length associated with transmitting the parameters can be estimated from

$$L(\boldsymbol{\theta}(i)) = \frac{1}{2} C \log\{B\} \tag{2.30}$$

where the number of free parameters in the model are denoted by C. The eigenvalues of the covariance matrix are real but the eigenvectors may be complex. Thus, it follows that there are $i + 2Mi + 1$ degrees of freedom resulting from these assertions. However, the eigenvectors are normalized to have unit length and are mutually orthogonal. Constraining the eigenvector length results in a loss of i degrees of freedom. Mutual orthogonality results in an additional loss of $2i(i-1)/2$ [45]. Accordingly, C has the value

$$C = i + 2Mi + 1 - i - i(i-1) = -i^2 + (2M+1)i + 1. \tag{2.31}$$

Combining the results from Eq. (2.29) and Eq. (2.30) with Eq. (2.22), the expected description length for a KL expansion model in additive Gaussian noise is obtained.

$$\begin{aligned} L(i) = -(M-i)B\Big(\log\Big\{\prod_{j=i+1}^{M} \lambda_{y_j}^{\frac{1}{M-i}}\Big\} - \log\Big\{\frac{1}{M-i}\sum_{j=i+1}^{M}\lambda_{y_j}\Big\}\Big) \\ + \frac{1}{2}(-i^2 + (2M+1)i + 1)\log\{B\} \end{aligned} \tag{2.32}$$

Finally, the MDL estimate of the model order K^* is determined by minimizing

$$K^* = \arg\min_i \mathrm{L}(i). \tag{2.33}$$

It has been shown in [46] that the MDL estimator has the following advantages:

- The estimated order permits the shortest encoding and captures all of the learnable data.
- The MDL criterion is consistent in the sense that it converges to the true model order as the sample size becomes large.

It should be noted that MDL was also discovered independently by Schwarz [47]. In his work, he showed MDL was optimal in the ML sense for minimizing error.

2.6.3 Merhav *et al.*'s Order Estimator

The order estimator proposed by Merhav *et al.* [48, 49] minimizes the underestimation probability while constraining the probability of overestimation to be exponentially decreasing with the order. Merhav *et al.* assumed that the signal has a Koopman-Darmois distribution function:

$$dP_{\boldsymbol{\theta}(i)} = \exp\{T(\boldsymbol{x})^T\boldsymbol{\theta}(i) - \kappa(\boldsymbol{\theta}(i))\}\mu(dx) \tag{2.34}$$

where $T(\cdot)$, $\mathbb{C}^M \rightarrow \mathbb{C}^K$, is a real valued sufficient statistic, i denotes the order of the model and μ is a σ-finite Lebesgue measure. $\kappa(\boldsymbol{\theta}(i))$ is defined to be the logarithmic moment generating function

$$\kappa(\boldsymbol{\theta}(i)) = \log\int_{-\infty}^{\infty}\exp\{T(\boldsymbol{x})^T\boldsymbol{\theta}(i)\}\mu(dx) \tag{2.35}$$

A Gaussian distribution may be obtained from the Koopman-Darmois distribution by an appropriate choice of parameters.

The optimization problem which Merhav *et al.* solve can be stated as

$$\begin{aligned} K^* &= \arg\min_i P_{\boldsymbol{\theta}(i)}(i < K) \\ \text{subject to: } &\lim_{n\to\infty}\left[-\frac{1}{n}\log P_{\boldsymbol{\theta}(i)}(i > K) > \beta\right] \end{aligned} \tag{2.36}$$

Assuming that a set of regularity conditions were satisfied, the optimal estimate of K can be obtained from

$$K^* = \arg\min_i \frac{1}{M}\log\frac{dP_{\boldsymbol{\theta}(K_0)}}{dP_{\boldsymbol{\theta}(i)}} < \beta \tag{2.37}$$

where K_0 is known to upper-bound the true order K". In this thesis, K_0 will be chosen to be the number of eigenvalues which exceed σ_w^2.

In the instance where the noise amplitude has a Gaussian distribution, the estimator of K can be calculated as [16]

$$K^* = \arg\min_{1\le i\le K_0}\left\{\frac{1}{2}\log\hat{\sigma}^2(i) - \frac{1}{2}\log\hat{\sigma}^2(K_0) < \beta\right\} \tag{2.38}$$

where $\hat{\sigma}^2(i)$ represents the energy of the noisy signal $\boldsymbol{y}$ in the noise subspace with dimension $M - K$.

$$\hat{\sigma}^2(i) = \frac{1}{M}\|\boldsymbol{Q_2}\boldsymbol{Q_2}^H\boldsymbol{y}\|_2^2 \qquad \boldsymbol{Q}_2 \in \mathbb{C}^{M\times(M-K)} \tag{2.39}$$

The estimate K^* is chosen such that the energy in the noise subspace does not increase beyond a factor of β times the upper bound of the energy.

The Merhav *et al.* estimator must be implemented differently from Ephraim and Van Trees in [16]. In their publication, β is chosen to be $0.0025\log\hat{\sigma}^2(K_0)$. Unfortunately, small noise variances may force β to be negative. As such, the upper bound of the order K_0 will always be chosen. It was experimentally determined that a fixed β for all values was more appropriate.

2.7 Linear Signal Estimators

Once the signal subspace has been determined, it remains to apply an estimator to the projection of the signal onto the signal subspace. Derivations for all estimators discussed in this section may be found in Appendix A.

The filter matrix for the linear estimator will be denoted as $\boldsymbol{H}$. The speech estimate resulting from applying the estimator can be calculated as $\hat{\boldsymbol{x}} = \boldsymbol{H}\boldsymbol{y}$. For the ensuing section, the residual of the clean signal, $\boldsymbol{e}$, can be represented as

$$\begin{aligned} \boldsymbol{e} &= \hat{\boldsymbol{x}} - \boldsymbol{x} \\ &= (\boldsymbol{H} - \boldsymbol{I})\boldsymbol{x} + \boldsymbol{w} \\ &\triangleq \boldsymbol{e}_x + \boldsymbol{e}_w \end{aligned} \tag{2.40}$$

$\boldsymbol{e}_x$ will be referred as the *signal distortion* while $\boldsymbol{e}_w$ will be denoted as the *residual noise.*

The energy of the signal distortion can be calculated from Eq. (2.41).

$$\epsilon_x^2 = \operatorname{tr} E\{\boldsymbol{e}_x\boldsymbol{e}_x^H\} = \operatorname{tr}\{(\boldsymbol{H} - \boldsymbol{I})\boldsymbol{R}_x(\boldsymbol{H} - \boldsymbol{I})^H\} \tag{2.41}$$

Similarly, the energy of the noise residual can be deduced from Eq. (2.42).

$$\epsilon_w^2 = \operatorname{tr} E\{\boldsymbol{e}_w\boldsymbol{e}_w^H\} = \sigma_w^2 \operatorname{tr}\{\boldsymbol{H}\boldsymbol{H}^H\} \tag{2.42}$$

Finally, the energy of the total error, ϵ can be calculated as

$$\epsilon^2 = \epsilon_x^2 + \epsilon_w^2 \tag{2.43}$$

2.7.1 Linear Least Square Estimator (LS)

The linear least square estimator minimizes the Euclidean distance between the noisy speech vector and the signal subspace.

$$\boldsymbol{H}^* = \arg\min_{\boldsymbol{H}} E\{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}\|_2^2\} \qquad \operatorname{rank}(\boldsymbol{H}) = K \tag{2.44}$$

It is easily shown that the filter that satisfies this problem is the orthogonal projector $\boldsymbol{Q}_1\boldsymbol{Q}_1^H$. This result should not be surprising since it was shown in Section 2.4.3 that the truncated KL expansion produces the optimal reduced-rank representation.

2.7.2 Minimum Variance Estimator (MV)

The minimum variance estimator attempts to minimize the mean-square error (MSE) of the estimated signal. This optimization can be expressed in the following manner:

$$\boldsymbol{H}^* = \arg\min_{\boldsymbol{H}} \epsilon_x^2 + \epsilon_w^2 \tag{2.45}$$

In this instance, the optimal filter is

$$\boldsymbol{H}^* = \boldsymbol{Q}_1(\Lambda_{x_1}(\Lambda_{x_1} + \sigma_w^2 \boldsymbol{I})^{-1})\boldsymbol{Q}_1^H \tag{2.46}$$

It should be noted that the derived filter closely resembles the frequency domain Wiener filter

$$\boldsymbol{H}_{\text{wiener}} = (\boldsymbol{S}_x(\boldsymbol{S}_x + \sigma_w^2 \boldsymbol{I})^{-1}) \tag{2.47}$$

where $\boldsymbol{S}_x = \text{diag}\{s_{x_1}, \ldots, s_{x_M}\}$ denotes the power spectral density matrix.

Modifications to the Minimum Variance Estimator

Ephraim and Van Trees showed in [16] that raising the coefficients of the eigendomain Wiener filter to the power γ decreased the audible residual noise while only slightly distorting the underlying speech signal. This estimator will be denoted as $\boldsymbol{H}_\gamma$.

$$\boldsymbol{H}_\gamma = \boldsymbol{Q}_1(\Lambda_{x_1}(\Lambda_{x_1} + \sigma_w^2 \boldsymbol{I})^{-1})^\gamma \boldsymbol{Q}_1^H \tag{2.48}$$

In addition, the generalized Wiener filter, as given in Eq. (2.49), was found to be effective for noise suppression.

$$\boldsymbol{H}_{\text{gen}} = \boldsymbol{Q}_1 \,\text{diag}(\exp\Big(-\frac{\sigma_w^2}{\lambda_{x_1}}\Big), \ldots, \exp\Big(-\frac{\sigma_w^2}{\lambda_{x_K}}\Big))\boldsymbol{Q}_1^H \tag{2.49}$$

It was also shown empirically that this estimator strongly attenuated the residual noise. However, the signal distortion increased accordingly. The name “Generalized Wiener Filter” derives from the relationship between the Wiener filter and the Taylor series of the inverse of $\exp\big(-\sigma_w^2/\lambda_{x_i}\big)$.

Observe that,

$$\exp\left(\sigma_w^2/\lambda_{x_i}\right) \approx 1 + \frac{\sigma_w^2}{\lambda_{x_i}} = \frac{\lambda_{x_i} + \sigma_w^2}{\lambda_{x_i}} \tag{2.50}$$

The MV filter is identical to the Spectral Domain Constraint proposed by Ephraim and Van Trees in [16]. However, the optimization problem has been reformulated to permit a more intuitive derivation.

2.7.3 Time Domain Constraint (TDC)

The Time Domain Constraint estimator minimizes signal distortion while constraining the average noise power to be less than $\alpha\sigma_w^2$. Thus,

$$\begin{gathered} \boldsymbol{H}^* = \arg\min_{\boldsymbol{H}} \epsilon_x^2 \\ \text{subject to: } \frac{1}{M}\epsilon_w^2 \leq \alpha\sigma_w^2 \end{gathered} \tag{2.51}$$

where $0 \leq \alpha \leq 1$. If α were not constrained to be less than one, the optimal solution would be the identity matrix $\boldsymbol{I}$. The resulting filter from the TDC constraints has the form

$$\boldsymbol{H}^* = \boldsymbol{R}_x(\boldsymbol{R}_x + \gamma\sigma_w^2\boldsymbol{I})^{-1} \tag{2.52}$$

where γ must satisfy

$$\alpha = \sigma_w^2 \operatorname{tr}(\boldsymbol{R}_x^2(\boldsymbol{R}_x + \gamma\sigma_w^2\boldsymbol{I})^{-2}) \tag{2.53}$$

2.8 Prewhitening Filter

If the additive noise $\boldsymbol{w}$ is not uncorrelated, it must be prewhitened. Prewhitening allows signal subspace methods to be applied to a much wider class of noise sources.

The signal is premultiplied by the matrix $\boldsymbol{R}_w^{-\frac{1}{2}}$ to decorrelate the noise signal.

$$\tilde{\boldsymbol{y}} = \boldsymbol{R}_w^{-\frac{1}{2}}\boldsymbol{y} = \boldsymbol{R}_w^{-\frac{1}{2}}\boldsymbol{x} + \boldsymbol{R}_w^{-\frac{1}{2}}\boldsymbol{w} = \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{w}} \tag{2.54}$$

It is then possible to derive the filter, $\tilde{\boldsymbol{H}}^*$ utilizing any of the linear estimators developed in Section 2.7.

$$\boldsymbol{H}^{*} = \boldsymbol{R}_{w}^{-\frac{1}{2}} \tilde{\boldsymbol{H}}^{*} \boldsymbol{R}_{w}^{\frac{1}{2}} \tag{2.55}$$

Filters designed using this method are not optimal as the values $\tilde{\epsilon}_x^2 = \mathrm{tr}\{\boldsymbol{R}_w^{\frac{1}{2}} E\{\boldsymbol{e}_x \boldsymbol{e}_x^H\} \boldsymbol{R}_w^{\frac{1}{2}}\}$ and $\tilde{\epsilon}_w^2 = \mathrm{tr}\{\boldsymbol{R}_w^{\frac{1}{2}} E\{\boldsymbol{e}_w \boldsymbol{e}_w^H\} \boldsymbol{R}_w^{\frac{1}{2}}\}$ are used to derive the estimators rather than ϵ_x^2 and ϵ_w^2.

2.9 Chapter Summary

This chapter discussed signal subspace enhancement systems. First, the idea of subspace decomposition was treated. It was then shown that signal subspace decomposition can be efficiently performed if a linear model is assumed and a decorrelative transform is utilized. The Karhunen-Loève transform was introduced and its advantageous properties were examined. Rank estimators for the signal under analysis were described. Then, linear estimators were outlined. These were utilized to obtain a better estimate once the signal subspace had been determined. Finally, methods to deal with coloured noise were discussed.

Chapter 3

Modelling Auditory Masking

The presence of one sound may interfere with the audibility of another. This auditory phenomenon is referred to as *masking*. Highly nonlinear in its nature, computationally intensive models are required to accurately describe the change in perception.

This chapter will introduce the concept of auditory masking. Subsequently, a physiological basis will be provided to explain the perceptual phenomenon. Afterwards, a mathematical model will be described to estimate the effects of masking. Knowledge of the masking phenomenon will be utilized later in Chapter 4 to produce a perceptual post-filter.

3.1 Masking - A Brief Introduction

Masking is the phenomenon where the perception of one sound is obscured by the perception of another [30]. A *masker* obscures a weaker signal known as the *maskee*. It is common to also refer to the maskee as the probe, target or signal. The threshold level above which a signal becomes audible in the presence of a masker is known as the *masking threshold* [50].

Masking effects occur when two sounds occur at the same time or when separated by a small delay. The former is known as simultaneous masking while the latter is known as temporal masking.

3.1.1 Measures of Audio Power

For psychoacoustics, the sound pressures between 10^{-5} Pa and 10^{2} Pa [51] are relevant. This implies that the ear has a dynamic range exceeding 140 dB.

Another commonly used measure is Sound Pressure Level (SPL). Sound pressure and SPL are related via the transformation

$$\mathrm{SPL} = 20 \log_{10} \left(\frac{p}{p_0} \right) \tag{3.1}$$

where p represents pressure (Pa) and, $p_0 = 20\ \mu\mathrm{Pa}$.

For signals with continuous spectra, it is common to refer to the energy, power or intensity contained within unit bandwidth. Such a measure has been termed as *noise power density*. The SPL may be calculated by integrating over the noise power density.

3.1.2 Simultaneous Masking

Simultaneous masking describes situations for which the masker is present during the entire duration of the maskee.

Classical experiments measuring masking usually involve narrow-band maskers and tones. It has been shown that for tonal maskers, the slope towards lower frequencies becomes steeper when increasing the magnitude of the masker. However, narrow-band noise maskers have been shown to have invariant low frequency slopes relative to masker level. For both classes of maskers, the curve towards higher frequencies become shallower with increasing SPL [52].

Furthermore, lower frequencies tend to mask higher frequencies. This phenomenon is known as the upward spread of masking. However, at levels below 40 dB, the reverse behaviour has been observed.

Moreover, when one compares a narrow band noise masker and a tonal masker of equal loudness, it can be shown that the noise masker is clearly the more effective masker. Hellman attributed these differences to the rate of intensity fluctuation inherent to narrow-band noise [53] that provides superior masking in comparison to tones.

3.1.3 Temporal Masking

Temporal masking has two typical forms: (1) forward masking and (2) backward masking.

Forward masking occurs when the maskee proceeds the masker. With a sufficiently short delay, the masker will obscure the maskee if it is within the same critical band (see Section 3.3.3 for details). In [54], it was shown that forward masking is effective up to 200 ms.

The prominence of forward masking is affected by three factors:

- The length of the masker.
- The delay between the masker and the maskee.
- The frequency of the masker and target.

Backward masking denotes the occasion where the maskee precedes the masker. In this instance, a loud sound will raise the threshold of hearing for a sound which has already occurred. A maximum delay of 20 ms is permissible for the effects of backward masking to be noticeable. The effect of backward masking is much weaker when compared to forward masking. Experiments have shown that practiced listeners may not even detect backward masking [55]. As such, it will not be considered for the remainder of this thesis.

3.1.4 Origins of Masking

Simultaneous masking has been linked to the process of swamping. Swamping can be understood as a masker producing sufficient activity in a critical band such that the underlying signal is no longer detectable [50].

The origins of forward masking are linked to several factors. They include:

- Temporal overlap of patterns of vibration on the basilar membrane (more important when the delay between masker and signal is small).
- An adaption process in the hair cells and the synapses between the hair cell and neurons in the auditory nerve [54].
- A persistence of the excitation evoked by the masker at a higher cognitive level than the auditory nerve.

It is thought that backward masking originates from blocking at a higher level than the cochlea [30].

3.2 Physiology of the Human Ear

The ear is comprised of three sections: outer, middle and inner ear. The outer ear predominantly has two functions: sound localization and funnelling sound into the ear. The middle ear essentially performs an impedance transformation between the low impedance air filled meatus and the high impedance cochlea. The inner ear processes speech signals prior to sending neural impulses to the brain. This analysis is carried out by the basilar membrane and the hair cells within the cochlea. A cross section of the ear is represented in Fig. 3.1

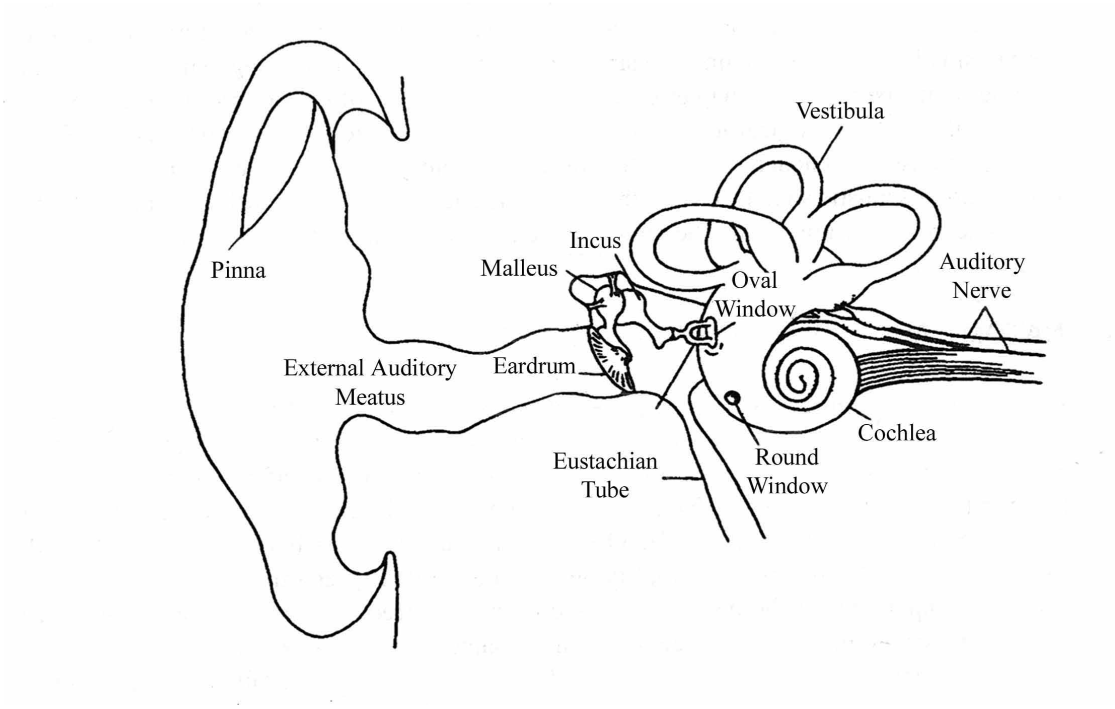


Fig. 3.1 Cross-sectional view of the hearing organs, from [30]

3.2.1 The Outer Ear

The outer ear consists of a partially cartilaginous flange called the pinna [56]. The main function of the outer ear is to direct incoming sound waves into the ear canal. Spectrally, the ear canal has a resonance at 2.5 kHz of 20 dB. Not coincidentally, this resonance occurs in the range of most energetic speech frequencies.

3.2.2 The Middle Ear

The eardrum marks the beginning of the middle ear. The sound is transmitted from the eardrum to three small bones, known as ossicles. The three ossicles are denoted as the malleus, incus and, stapes (also known as the hammer, anvil and stirrup). The stapes is attached to the oval window, the interface between the middle ear and cochlea. The middle ear acts as a low pass filter with attenuation of -15 dB/octave above 1 kHz.

The middle ear performs an impedance transformation from a large, low impedance tympanic membrane to the much smaller high impedance oval window. The input impedance of the cochlea was measured by Lynch *et al.* [57] to be 1.5×10^5 N sec/m^3. This is in stark contrast with air which has an impedance of 430 N sec/m^3. Without such a transformation, only 0.1% of the pressure waves hitting the eardrum would be transmitted to the inner ear [30].

3.2.3 The Inner Ear

The inner ear is responsible for transforming the mechanical pressure waves that are incident on the oval window into neural impulses. This is accomplished by the cochlea and the neural hairs that lie within it.

The Cochlea

The cochlea has dimensions of 1 cm wide and 5 mm from base to apex. The uncoiled cochlea has a length of 35 mm. A cross-sectional view of the cochlea is shown in Fig. 3.2.

The cochlea can divided into three distinct scalae (chambers). These three scalae are known as vestibuli, media and, tympani. The scala vestibuli is separated from the scala media by Reissner's membrane. The basilar membrane divides the scala media from the scala tympani. Vibrations of the oval window are transmitted to the perilymph contained in the scala vestibuli. This aperture is located at the base of the cochlea. The heliocotrema allows the release of pressure from the scala vestibuli to the scala tympani. The round window provides a similar service in the scala tympani. As the walls of the cochlea are hard bone, and the liquid is incompressible, motion of the oval window produces vibration in the membranes of the cochlea.

The scala media is filled with endolymph, a gelatinous liquid. Inside, it contains the

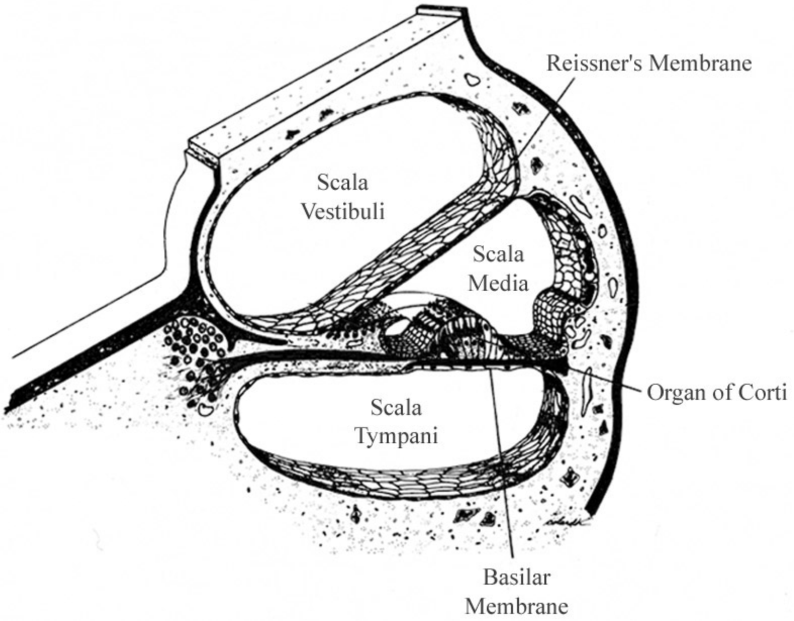


Fig. 3.2 Cross-sectional view of the cochlea, from [56]

basilar membrane. Within this structure lies the basilar membrane, a fibrous, elastic structure which makes 2.75 turns as it follows the walls of the cochlea [56].

The scala media is broad near the base and tapers off towards the apex. This is in direct opposition to the basilar membrane which is narrow at the base and wide at the apex.

Each point along the basilar membrane has a characteristic frequency to which it vibrates most. The basilar membrane is stiff at the basal end, yet compliant at the apex. Accordingly, the apex responds to lower frequency stimuli while the base maximally vibrates for high frequency sensations.

Neural Hairs

The hair cells are arranged in four or five rows along the basilar membrane. The hair cells are connected to the brain via the eighth cranial nerve (auditory nerve). This conduit is composed of neurons which fire in response to the bending and shearing forces resulting from the motion of the basilar membrane.

Insight into the behaviour of hair cells can be obtained by studying neurological tuning curves. A neurological tuning curve depicts the intensity of a sinusoidal tone required for a neurone to fire at a preselected rate. An example of this measurement for cats is displayed in Fig. 3.3.

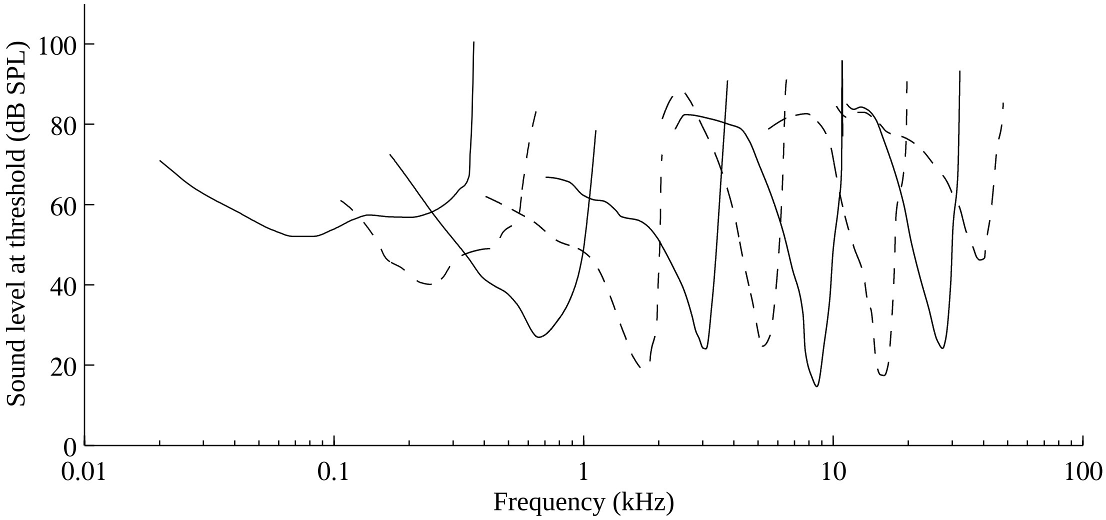


Fig. 3.3 Neurological tuning curves of cats, from [50]

It becomes evident that the neural hairs display behaviour consistent with the mechanical motion of the basilar membrane. Further, the tuning curves resemble constant Q bandpass filters. Thus, the ear has greater resolution at lower frequencies, and poorer resolution at higher frequencies.

3.3 Calculation of the Masking Threshold

Typically masking models attempt to model the physiological elements of the ear. Accordingly most frequency domain models have the following subsystems:

- Windowing, Normalization and Fourier Transform of Data: This stage buffers the masker into overlapping frames of data. A window is applied to improve the spectral properties of the frame of data. Afterwards, the quantized signal is mapped to SPL and converted to the frequency domain.

- Outer/Middle Ear Transfer Function: The filtering performed by the ear prior to the oval window is represented by this component.

- Critical Band Integration: This system models the frequency resolution of the ear. The energy over a critical band[1] is summated to mimic the frequency discrimination abilities of the human auditory system.

- Addition of Internal Noise: The body contributes stimulation which the ear detects. As a result, a frequency dependent offset is added to simulate internal noise within the auditory system.

- Energy Spreading: The firing patterns of the neural hairs contained within the cochlea similarly to a set of bandpass filters. This functional block models the spectral behaviour of this physiological component.

- Temporal Masking: The stimulation from past maskers affects the present masking threshold. This subsystem incorporates data from past frames to model this effect.

- Masking Offset: The masking threshold is calculated by applying a frequency dependent offset onto the excitation pattern.

The masking model utilized in this work was described in the ITU perceptual quality assessment tool (PEAQ) standard [58]. Originally designed to operate with a sampling rate of 48 kHz and a window length of 40 ms, the masking model has been adapted to operate at other sampling rates and window sizes.

3.3.1 Windowing, Normalization and Fourier Transform of Masker Signal

The masker must be windowed before further processing can take place. The window size, denoted as N, must be chosen carefully such that there is sufficient frequency resolution. However, if the window becomes too large, the effects of temporal masking will not be well resolved.

Afterwards, the speech signal must be converted from its quantized representation to SPL. This is done by utilizing a normalization constant. The value is determined by analyzing a full-scale 1019.5 Hz tone over 10 frames. This tone represents the maximum noise power density attainable and will be normalized to be 92 dB. Clearly, this computation need only be performed once.

[1] A critical band describes the resolution of the human ear of a particular frequency.

Finally, the Fourier transform of the signal is taken and normalized by the length of the data window.

3.3.2 Outer/Middle Ear Transfer Function

The outer/middle ear transfer function models the spectral properties of the pinna and auditory canal. The resonances of these structures are discussed in Section 3.2.

PEAQ Model of The Outer/Middle Ear Transfer Function

The spectral behaviour of the outer and middle ear is modelled in the ITU standard as [58]

$$W[i] = -0.6 \cdot 3.64 \Big(\frac{f[i]}{1000}\Big)^{-0.8} + 6.5 \cdot \exp\Big(-0.6 \cdot \Big(\frac{f[i]}{1000} - 3.3\Big)^2\Big) - 10^{-3} \cdot \Big(\frac{f[i]}{1000}\Big)^{3.6} \text{dB} \quad (3.2)$$

where $i \in [0, N-1]$. This model was based on the work of Terhardt in [59].

3.3.3 Critical Band Grouping

Critical bands represent the frequency resolution of the ear. Within a critical band, the spectral structure of the masker is unimportant. Rather, the overall energy of the masker affects perception.

Description of Critical Bands

Fletcher first described critical bands in [60]. He did so by measuring the masking threshold of a sinusoidal signal in the presence of a noise masker with changing bandwidth. Fletcher held the power density of the noise masker constant and centred it about the sinusoidal tone. It was discovered that the masking threshold increased with the bandwidth of the noise signal until the bandwidth exceeded a fixed amount. After which, the SPL required for the perception of the tone did not increase. This threshold was known as a critical bandwidth.

Critical bands correspond to approximately 1.5 mm spacings along the basilar membrane. As a reasonable approximation to the true nature of continuous critical bands, many masking models divide the frequency range under analysis into non-overlapping critical bands.

PEAQ Model of Critical Bands

The *Bark* is a perceptual scale which relates the concept of non-overlapping filters to frequency. One Bark spans the width of a critical band. Schroeder *et al.* [61] approximated this scale using the function

$$z = 7 \cdot \operatorname{arcsinh}\left(\frac{f}{650}\right). \tag{3.3}$$

For added resolution PEAQ utilizes quarter-bark groupings. This ensures better resolution of the masking threshold estimate at higher frequencies. The total number of quarter-bark bands will be denoted as Z. For the frequency range of 0–4000 Hz, $Z = 69$.

Once this is done, the PEAQ model calculates the energy in each of the quarter-bark bands. If a frequency bin straddles two bands, the energy is multiplied by the percentage of the frequency bin lying within the critical band. The resulting values, $P_e[k]$, are denoted as the *energies of the frequency groupings.* It should be noted that $k \in [0, Z-1]$. The grouping algorithm is presented in its entirety in Appendix B.

3.3.4 Addition of Internal Noise

Heart beats and spontaneous activity of muscles produce internal noise [59]. This stimulation source is responsible for the rise in threshold in quiet at low frequencies. The internal noise is frequency independent at medium and high frequencies, but rises strongly at low frequencies.

PEAQ Model of Internal Noise

The effect of internal noise is modelled as [58]

$$P_{\text{int}}[k] = 10^{0.4 \cdot 3.64 \left(\frac{f[k]}{1000}\right)^{-0.8}}. \tag{3.4}$$

The masker power calculated in Section 3.3.3 and the internal noise power are added linearly to obtain P_p, the *pitch patterns.*

$$P_p[k,n] = P_e[k,n] + P_{\text{int}}[k] \tag{3.5}$$

Auditory Threshold of Hearing

Before a sound can be detected, it must exceed the *auditory threshold of hearing*. This is the minimum intensity level at which sounds are detectable. The shape of the auditory threshold results from the filtering effects of the outer and middle ear. It can also be attributed to the internal noise of the body.

PEAQ modelled the auditory threshold of hearing over two blocks: the *outer/middle ear transfer function* and the *addition internal noise*. A plot of the threshold of hearing is given Fig. 3.4

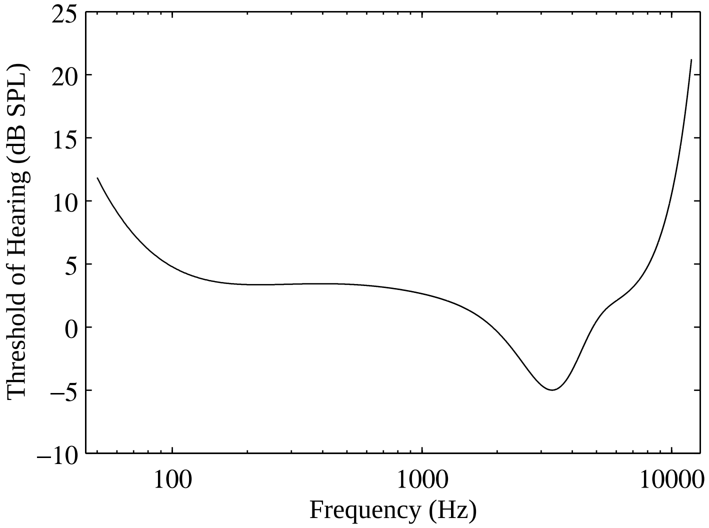


Fig. 3.4 Auditory threshold of hearing, from [28]

The auditory threshold remains relatively constant for the speech frequencies from 700 Hz and 7000 Hz with the hearing threshold stays within ± 3 dB [30].

3.3.5 Energy Spreading

The neural activity resulting from the current frame, n, stimulus is estimated using spreading functions and specialized additivity laws.

Spreading Function

Most masking models attempt to model the filter-like behaviour of the neural hairs by convolving the critical band densities with a spreading function. Terhardt proposed the following model in [62].

$$\begin{aligned} S_l[k, L[k,n]] &= 27 \\ S_u[k, L[k,n]] &= -24 - \frac{230}{f_c[k]} + 0.2L[k,n] \end{aligned} \tag{3.6}$$

where $L[k,n] = 10\log_{10}(P_p[k,n])$. The lower slope of the spreading functions is assumed to be independent of SPL and frequency. Conversely, the upper slope is very much dependent on SPL and weakly affected by frequency.

Additivity of Simultaneous Maskers (Power Law)

To model the nonlinear additivity of maskers, many models have been proposed. Two maskers together produce substantially more masking than either masker produces separately. This was exemplified by an experiment performed by Green [63]. He measured the combined effect of two maskers, a broad band noise and a sinusoid, with equal masking at their centre frequency. If linear addition were appropriate, a 3 dB increase should have been visible. However, 6 to 14 dB of additional masking was detected. This phenomenon has referred to as *excess masking.*

Lufti proposed the *power law* to better account for the additional masking [64]. The additivity of simultaneous can be modelled as

$$X_{ab}^p = X_a^p + X_b^p. \tag{3.7}$$

PEAQ Model of Energy Spreading

The PEAQ model utilizes a factor of $p = 0.4$ in the *power-law* to model the additivity of maskers. The Schroeder *et al.* spreading function is employed to model the frequency selectivity of the basilar membrane. This operation produces the *unsmeared excitation* $E_2[k,n]$. The convolution has the form

$$E_2[k,n] = \frac{1}{\mathrm{Norm}[k]} \Big(\sum_{j=0}^{Z} E_{\mathrm{line}}[j,k,n]^{0.4} \Big)^{\frac{1}{0.4}}. \tag{3.8}$$

Eq. (3.8) contains two forms of normalization. These account for increases of signal energy from the spreading function. Firstly, it ensures that the energy of the spreading function is 0 dB. This is accomplished by the summation in the denominator of Eq. (3.9).

$$E_{\text{line}}[j,k,n] = \begin{cases} \dfrac{10^{\frac{L[j,n]}{10}}\, 10^{\frac{-0.25(j-k)S_l[j,L[j,n]]}{10}}}{\displaystyle\sum_{\mu=0}^{j-1} 10^{\frac{-0.25(j-\mu)S_l[j,L[j,n]]}{10}} + \sum_{\mu=j}^{Z-1} 10^{\frac{-0.25(\mu-j)S_u[j,L[j,n]]}{10}}} & k < j \\[2em] \dfrac{10^{\frac{L[j,n]}{10}}\, 10^{\frac{0.25(k-j)S_u[j,L[j,n]]}{10}}}{\displaystyle\sum_{\mu=0}^{j-1} 10^{\frac{-0.25(j-\mu)S_l[j,L[j,n]]}{10}} + \sum_{\mu=j}^{Z-1} 10^{\frac{-0.25(\mu-j)S_u[j,L[j,n]]}{10}}} & k \geq j \end{cases} \tag{3.9}$$

Additionally, it is ensured that if a signal with a constant 0 dB noise power density is presented to the algorithm, an unsmeared excitation of 0 dB will be returned. Originally proposed by Johnston in [65], this operation approximates the deconvolution of the spreading function and the unsmeared excitation. This normalization is carried out by Norm$[k]$ term where

$$\text{Norm}[k] = \left(\sum_{j=0}^{Z} \tilde{E}_{\text{line}}[j,k]^{0.4}\right)^{\frac{1}{0.4}} \tag{3.10}$$

and

$$\tilde{E}_{\text{line}}[j,n] = \begin{cases} \dfrac{10^{\frac{-0.25(j-k)S_l[j,0]}{10}}}{\displaystyle\sum_{\mu=0}^{j-1} 10^{\frac{-0.25(j-\mu)S_l[j,0]}{10}} + \sum_{\mu=j}^{Z-1} 10^{\frac{-0.25(\mu-j)S_u[j,0]}{10}}} & k < j \\[2em] \dfrac{10^{\frac{0.25(k-j)S_u[j,0]}{10}}}{\displaystyle\sum_{\mu=0}^{j-1} 10^{\frac{-0.25(j-\mu)S_l[j,0]}{10}} + \sum_{\mu=j}^{Z-1} 10^{\frac{-0.25(\mu-j)S_u[j,0]}{10}}} & k \geq j \end{cases} . \tag{3.11}$$

3.3.6 Forward Masking

The excitation pattern describes the resulting neural activity evoked by a sound as a function of the characteristic frequency of the neurones being excited.

Forward masking in PEAQ is modelled by an envelope detector. This device has two functional components: a smoothing filter and a maximization function.

The smoothing function in PEAQ is modelled by a first order IIR filter defined by the difference equation

$$E_f[k,n] = a[k]E_f[k,n-1] + (1-a[k])E_2[k,n] \tag{3.12}$$

where $a[k] = \exp(-T_d/\tau[k])$ and T_d represents the frame advancement after each iteration. The time constant $\tau[k]$ has been chosen to be

$$\tau[k] = \tau_{\min} + \frac{100}{f_c[k]}(\tau_{100} - \tau_{\min})\Bigg|_{\substack{\tau_{100}=0.030\\ \tau_{\min}=0.008}}. \tag{3.13}$$

By changing the time constant as a function of k, the dependency on frequency of forward masking is well-modelled. To handle fast attacks, the following maximization function was used

$$E[k,n] = \max\{E_f[k,n], E_2[k,n]\}. \tag{3.14}$$

This function allows rapid rises in noise power density when the masker increases in amplitude. However, drops in noise power density result in a slow decay of excitation due to the low-pass filtering of Eq. (3.13). $E[k,n]$ is known as the excitation pattern.

3.3.7 Subtraction of Masking Offset

To determine the masking threshold, a weighting function is applied to the excitation pattern calculated in the previous section. The weighting function has been defined as [58]

$$z[k,n] = \begin{cases} 3 & k \leq 48 \\ (0.25)^2 k & k > 48 \end{cases}. \tag{3.15}$$

Finally, the masking threshold, $m[k,n]$, is obtained

$$m[k,n] = \frac{E[k,n]}{10^{\frac{z[k,n]}{10}}}. \tag{3.16}$$

It should be emphasized that m[k,n] is defined with respect to the inner ear using the Bark scale. The masking threshold denotes the minimum intensity a maskee must have to be perceived in the presence of a masker.

3.4 Chapter Summary

The auditory phenomena of masking was introduced in this chapter. Its origin and characteristics were described in detail. The physiology of the human ear was also discussed. Finally, the PEAQ masking model was presented in its entirety. The individual steps of the algorithm were related to the auditory mechanisms which they mimicked.

Chapter 4

Design of a Perceptual Post-Filter

The subspace filter described in Chapter 2 has been shown to be effective in improving the Signal-to-Noise Ratio (SNR) of an speech signal [16]. Though, this method has also been found to introduce artefacts into the enhanced signal. These artefacts are known as *musical noise* and have often been evaluated as being more disturbing than the original corrupting noise.

To remove these annoyances, a perceptual post-filter will be employed. This system will smooth the output of the signal subspace filter and reduce the prominence of the musical noise. By utilizing properties of the human auditory system, the underlying speech signal should remain largely undistorted.

This chapter will describe the phenomenon of musical noise. It will be shown that the incorporation of the principle of masking into an auditory post-filter will reduce these audible artefacts. Finally, an algorithm based on signal subspace methods utilizing an auditory post-filter will be outlined.

4.1 Musical Noise

Musical noise is a major problem in speech enhancement. Several speech enhancement schemes have attempted to address the problem using various approaches. These have included time averaging [12, 66], noise floors, oversubtraction of noise [13] or perceptual criteria [28, 29].

Musical noise is characterized from randomly spaced peaks in the spectrum of the reconstructed signal. Transformed back to the time domain, they will sound similar to the sum of tone generators with random fundamental frequencies which are rapidly turned on and off [12].

A primary source of musical noise is the crude estimation of the noisy signal power spectrum that is used in most speech enhancement algorithms. The use of raw input spectral amplitude tends to produce inaccurate filter magnitude response owing to the random noise and signal fluctuations within each frequency band. This results in over-estimates and under-estimates of the clean signal in adjacent spectral groupings.

Signal subspace techniques prevent musical noise from this origin by calculating the noisy signal correlation matrix, $\boldsymbol{R}_y$, over long windows and processing speech over very short frames. Accordingly, the eigenvectors used to provide a basis for the signal subspace will vary slowly from frame-to-frame. Thus, they are very unlikely to cause musical noise.

Unfortunately, signal subspace techniques introduce musical noise in two other manners. Firstly, sudden changes in estimated model order can produce fluctuating tone-like components. In addition, new peaks may be suddenly added if signal subspace eigenvectors are confused with noise subspace eigenvectors. This phenomenon is known as *subspace swapping* [67, 68].

An example of musical noise has been provided in Fig. 4.1. This plot was created by the application of the *generalized spectral subtraction algorithm* to a signal of pure white noise. A detailed description of this algorithm can be found in Section 4.6.1. It can be seen that while low amplitude noise has been attenuated to zero as desired, the high amplitude components have remained largely unsuppressed, owing to the nonlinear nature of the suppression rule [66].

4.2 Motivation for the Perceptual Post-Filter

It is the goal of the perceptual post-filter to remove all traces of musical noise. Its strengths are two-fold: (1) distortion is minimized by attenuating only what is audible, and (2) peaks within the noise residual are smoothed by spectral and temporal averaging. However, the underlying speech should not be affected. Such systems have been used successfully in [28, 29, 69] for speech enhancement.

Limiting the attenuation in an enhancement scheme can decrease the production of artefacts. Perceptual filters accomplish this by suppressing until the residual noise lies below the masking threshold. As such, some noise which is imperceptible is retained. By

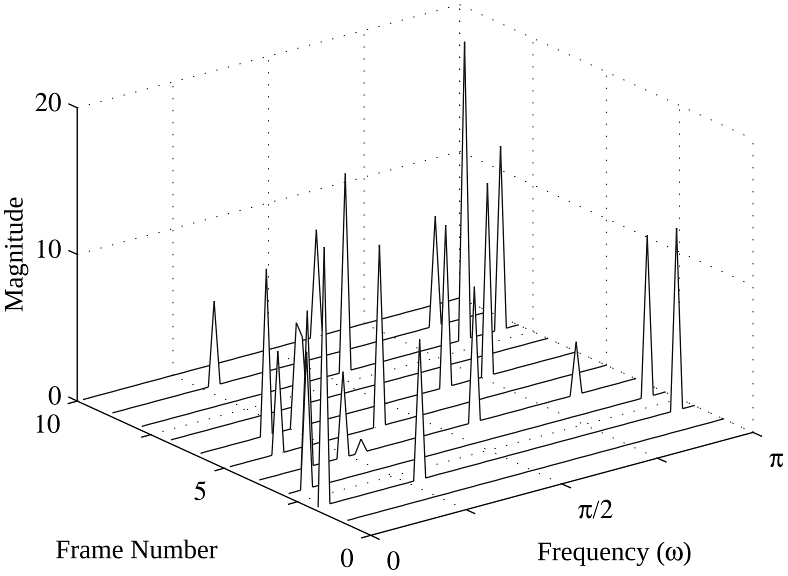


Fig. 4.1 Example of musical noise

attenuating less, it is expected that fewer disturbances will be produced. For the listener, there should not be a discernible increase in residual noise as compared with conventional algorithms.

By considering human perception, artefacts can be smoothed without noticeably altering the underlying speech signal. Spectral averaging increases the width of tones within the noise residual according to the resolution of the ear. Rapid frame-to-frame spectrum variations are with high probability, the product of noise. Temporal averaging, by limiting magnitude changes of the noise residual over several frames, effectively attenuates musical noise.

4.3 Overview of the Enhanced Signal Subspace Method

An overview of the Enhanced Signal Subspace (ESS) method will be given presently. The signal subspace filter will be modified to suppress musical noise by appending a perceptual post-filter. A flow-chart describing the operation of the modified speech enhancement scheme as shown in Fig. 4.2. A more extensive discussion regarding the functionality of the individual blocks will follow in subsequent sections.

The signal subspace filter operates most effectively when utilizing very short frames (<

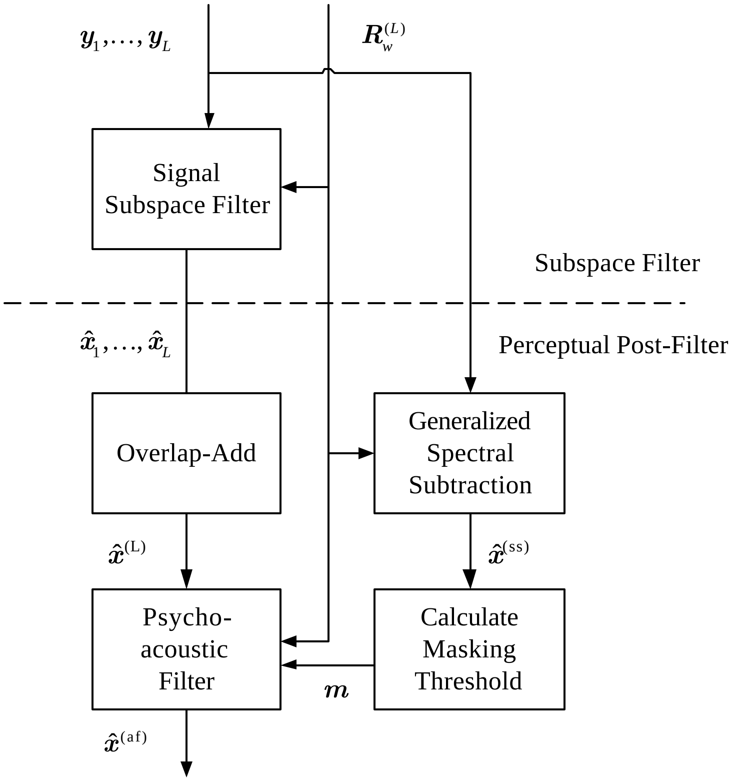


Fig. 4.2 Flow chart of the improved signal subspace enhancement scheme

15 ms). Unfortunately, such frames do not provide sufficient frequency resolution for the calculation of a masking threshold. Thus, L input frames are sent to the signal subspace filter ($\boldsymbol{y}_1,\ldots,\boldsymbol{y}_L$). The outputs, $\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_L$, are later merged, thereby increasing the frequency resolution of the estimate. The labels $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_L$ will denote the noise in the L input frames.

The psychoacoustic filter attempts to conceal the salient noise using the perceptual properties of the ear while minimizing the distortion to the underlying speech. This block is signal dependent, requiring an estimate of the noise correlation matrix, $\boldsymbol{R}_w^{(L)}$ and the masking threshold of the speech signal, $\boldsymbol{m}$, to calculate an appropriate gain.

The input to the psychoacoustic filter is $\hat{\boldsymbol{x}}^{(L)}$, the concatenation of L output frames from the signal subspace filter, $\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_L$. The frames are combined by the overlap-add block that utilizes appropriate windows and overlap length. The symbol $\boldsymbol{w}^{(L)}$ will denote the concatenated noise frames ($\boldsymbol{w}_1,\ldots,\boldsymbol{w}_L$).

The masking threshold is calculated utilizing the model described in the Perceived

Audio Quality ITU Recommendation (ITU-R BS.1387) [58]. However, the masking model has been modified to operate with a range of sampling frequencies and time windows. As the clean speech signal is unavailable, it is necessary to estimate the masking threshold of the speech signal from noisy data. The spectra of the clean speech is approximated using the spectral subtraction technique.

4.4 Description of the Psychoacoustic Filter

The psychoacoustic filter eliminates audible noise using a perceptual criterion. It is designed in the frequency domain to allow the vast sums of knowledge related to auditory perception to be applied.

The matrix representation of the discrete Fourier transform will be first, reviewed to allow its use in the definition of the constrained optimization problem.

Definition 2 (Discrete Fourier Transform) *The matrix $\boldsymbol{F}$ will represent the Discrete Fourier transform matrix where*

$$f_{n,k} = \exp\Big(-\frac{j2\pi kn}{N}\Big). \tag{4.1}$$

Accordingly the DFT of a vector $\boldsymbol{u}$ is defined as

$$\mathcal{F}\{\boldsymbol{u}\} = \boldsymbol{F}^H \boldsymbol{u}. \tag{4.2}$$

The auditory filter minimizes signal distortion while constraining the spectrum of the noise residual to be beneath the masking threshold. The constrained optimization problem can be summarized as

$$\begin{gathered}\boldsymbol{T}^* = \arg\min_{\boldsymbol{T}} \; E\{\|(\boldsymbol{T}\boldsymbol{F}^H - \boldsymbol{F}^H)\hat{\boldsymbol{x}}^{(L)}\|_2^2\}] \\ \text{subject to: } E\{|\boldsymbol{f}_i^H \boldsymbol{T}\boldsymbol{w}^{(L)}|^2\} \le m_i^2 \end{gathered} \tag{4.3}$$

where $\boldsymbol{T}^* = \text{diag}\{t_1^*, \ldots, t_N^*\}$. N denotes the concatenated frame size and $\boldsymbol{F}$ is the Fourier transform matrix.

A derivation for a closed-form solution for this optimization problem is given in Appendix A. The optimal filter can be formulated as

$$t_i^* = \begin{cases} 1 & m_i \geq (s_{w_i}^{(L)})^{\frac{1}{2}} \\ \dfrac{m_i}{(s_{w_i}^{(L)})^{\frac{1}{2}}} & m_i < (s_{w_i}^{(L)})^{\frac{1}{2}} \end{cases} . \tag{4.4}$$

where $\boldsymbol{s}_w^{(L)}$ is the power spectral density of the composite residual noise.

Intuitively, it may seem more appropriate to attempt to place the noise residual of the signal subspace filter beneath the masking threshold. However, it was determined empirically that this criteria added musical noise due to the peakiness of the filter transfer function.

4.5 Description of Overlap-Add Operation

Several frames must be combined to produce sufficient information for the creation of the psychoacoustic filter. This is accomplished by superimposing several small frames with an overlap of P. Henceforth, the smaller frames produced by the signal subspace filter will be denoted as *subframes*. The combined subframes will be referred to simply as *frames*. It is necessary to introduce overlap to minimize block edge effects. The overlapped frames are multiplied by a window and summed. The window ensures that there are smooth transitions at the block edges. This operation is summarized in Eq. (4.5).

$$\hat{x}_n^{(L)} = \sum_{i=0}^{L-1} x_n^{(i)} w_n^{(i)} \qquad n \in [0, N-1] \tag{4.5}$$

4.5.1 Selection of the Overlap-Add Window

The window used for the overlap-add operation must satisfy the property

$$\sum_{i=0}^{L-1} w_n^{(i)} = 1 \qquad n \in [0, N-1]. \tag{4.6}$$

Two different classes of windows are utilized to produce the combined frames. The initial and final subframes are multiplied by a window with a smooth rolloff on only one extremity. All other subframes are multiplied by a window with rolloff on both extremities.

$$w_i^{(0)} = \begin{cases} 1 & 0 \leq i \leq M-P-1 \\ \sin^2\left\{\frac{\pi(M-1-i)}{2(P-1)}\right\} & M-P \leq i \leq M-1 \\ 0 & \text{otherwise} \end{cases} \tag{4.7a}$$

$$w_i^{(l)} = \begin{cases} \sin^2\left\{\frac{\pi(i-nM)}{2(P-1)}\right\} & nM \leq i \leq nM+P-1 \\ 1 & nM+P \leq i \leq (n+1)M-P-1 \\ \sin^2\left\{\frac{\pi(M-1-(i-nM))}{2(P-1)}\right\} & (n+1)M-P \leq i \leq (n+1)M-1 \\ 0 & \text{otherwise} \end{cases} \tag{4.7b}$$

$$w_i^{(L-1)} = w_{N-i}^{(0)} \tag{4.7c}$$

where M denotes the size of the subframe and $l \in [1, L-2]$.

4.5.2 Calculation of the Noise Power Spectral Density

The Blackman-Tukey spectrum estimation method will be employed to estimate the power spectral density of the noise signal [70]. It makes use of the Wiener-Khinchin theorem which dictates that the relationship between the autocorrelation sequence and the power spectral density of a Wide-Sense Stationary (WSS) random process.

Definition 3 (Wiener-Khinchin Theorem) *The power spectral density of x_n, a zero mean random variable, is defined as the Fourier transform of the autocorrelation sequence*

$$S_x(f) = \mathcal{F}\{r_x\} \tag{4.8}$$

The power spectrum is obtained by calculating the Fourier transform of the windowed autocorrelation sequence. The autocorrelation sequence is windowed to deemphasize the largest lags. The correlation estimates of these lags have a large variance due to the smaller number of samples being used to determine their value.

The autocorrelation sequence has been chosen to be of length D where

$$D = \begin{cases} N+1 & N \text{ is even} \\ N & N \text{ is odd} \end{cases}. \tag{4.9}$$

It is necessary that the window be of odd length and symmetric to force the power spectral density to be real. Furthermore, it is desirable that the window spectrum be nonnegative. This will ensure that the derived power spectral density is also nonnegative. For this implementation, the window applied to the autocorrelation sequence will be a Bartlett window

$$w_b = \begin{cases} 1 - |i|/\frac{D-1}{2} & i < \frac{D-1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

The power spectral density samples are calculated by taking a D-point DFT while sampling the spectra with only N points. Thus, the Fourier transform will be redefined as

$$\hat{f}_{n,k} = \exp\Bigl(-\frac{j2\pi kn}{N}\Bigr) \tag{4.11}$$

where $\hat{\boldsymbol{F}} \in \mathbb{C}^{D\times N}$.

In the time domain, this operation is equivalent to folding the autocorrelation sequence with a period with N rather than D. Thus, if N is even, there is aliasing with the largest lag. However, this lag tends to be quite small, minimizing any ill-effects.

Thus, the Blackman-Tukey estimate of the power spectral density can be found as

$$\boldsymbol{s}_x^{(L)} = \hat{\boldsymbol{F}}^H \boldsymbol{W}_b \boldsymbol{r}_x \tag{4.12}$$

where $\boldsymbol{W}_b = \operatorname{diag}\{w_{b_1}, \ldots, w_{b_D}\}$.

4.6 Estimation of Masking Threshold From Noisy Data

The PEAQ model is utilized to model auditory perception. However, as the clean speech signal is unavailable, it is necessary to estimate the masking threshold of the speech signal from noisy data. The generalized spectral subtraction algorithm is employed to obtain a crude estimate of the speech signal. From this the masking threshold is calculated.

Afterwards, it is necessary to modify the masking threshold such that it is given with respect to the sound apparatus of the outer ear in the frequency scale.

4.6.1 Generalized Spectral Subtraction Algorithm

An estimate of the clean speech signal is required for an accurate masking threshold. This coarse approximation will be obtained from the generalized spectral subtraction algorithm. Spectral subtraction is based on the relationship given in Eq. (4.13) for signals corrupted by uncorrelated noise

$$\boldsymbol{s}_y = \boldsymbol{s}_x + \boldsymbol{s}_w. \tag{4.13}$$

Clearly, the magnitude response of the speech signal can be estimated from power subtraction. The noisy phase is retained in the enhancement system. As the masking threshold is insensitive to phase, this approximation should not affect the performance of the perceptual post-filter.

The generalized spectral subtraction algorithm was defined by Virag in [29]. It is a hybrid of the spectral subtraction algorithm proposed by Berouti *et al.* [13] and the generalized noise reduction outlined by Lim and Oppenheim in [71].

If it can be assumed that the additive noise is slowly changing, an adequate estimate of $\boldsymbol{s}_w^{(L)}$ can be obtained by averaging the magnitude response of $\mathcal{F}\{\boldsymbol{w}\}$ over several frames. Unfortunately, similar averaging of the magnitude response of $\mathcal{F}\{\boldsymbol{y}\}$ cannot be performed. Otherwise, sudden onsets will be significantly "dulled" due to the low-pass filtering of the time trajectories of the DFT coefficients. Thus, the instantaneous value $|\boldsymbol{F}^H\boldsymbol{y}|$ is utilized.

The generalized spectral subtraction algorithm incorporates oversubtraction and spectral flooring to minimize musical noise.

Oversubtraction removes more noise than necessary. As such, the narrow peaks which create musical noise are removed. Unfortunately, the reduction of musical noise is achieved at the cost of speech signal distortion.

Spectral flooring techniques attempt to reduce musical noise by filling spectral valleys with noise. The sharpness of the peaks is reduced and the amount of musical noise perceived is diminished. Alternatively, it can be thought that the noise floor masks the musical noise [13].

The gain function of the spectral function is defined uniquely by a posteriori SNR

$\dfrac{|\boldsymbol{F}^H\boldsymbol{y}|_i}{|\boldsymbol{F}^H\boldsymbol{w}|_i}$.

$$g_i\left[\frac{|\boldsymbol{F}^H\boldsymbol{y}|_i}{|\boldsymbol{F}^H\boldsymbol{w}|_i}\right] = \begin{cases} \left(1-\alpha\left[\dfrac{|\boldsymbol{F}^H\boldsymbol{w}|_i}{|\boldsymbol{F}^H\boldsymbol{y}|_i}\right]^{\gamma}\right)^{\frac{1}{\gamma}} & \left[\dfrac{|\boldsymbol{F}^H\boldsymbol{w}|_i}{|\boldsymbol{F}^H\boldsymbol{y}|_i}\right]^{\gamma} < \dfrac{1}{\alpha+\beta} \\ \left(\beta\left[\dfrac{|\boldsymbol{F}^H\boldsymbol{w}|_i}{|\boldsymbol{F}^H\boldsymbol{y}|_i}\right]^{\gamma}\right)^{\frac{1}{\gamma}} & \left[\dfrac{|\boldsymbol{F}^H\boldsymbol{w}|_i}{|\boldsymbol{F}^H\boldsymbol{y}|_i}\right]^{\gamma} \geq \dfrac{1}{\alpha+\beta} \end{cases} . \tag{4.14}$$

In Eq. (4.14), α represents the oversubtraction factor. Typically, α would have a value greater than 1.

The parameter β controls the spectral noise floor. It determines the minimum value that the gain function in Eq. (4.14) may assume [29].

Finally, γ determines the sharpness of transition from $g_i[-\infty] = 0$ to $g_i[\infty] = 1$. Usually, $\gamma \geq 1$ with increasing γ producing more attenuation.

The parameters α, β and γ allow one to increase attenuation to minimize musical noise. It was determined empirically that if the estimate contains musical noise, the output of psychoacoustic filter will contain a similar structure. Sharp peaks must be attenuated to provide a smooth masking threshold.

4.6.2 Normalization of the Masking Threshold

The masking threshold derived from the PEAQ algorithm must be modified in two different ways to operate successfully with the psychoacoustic filter. The masking threshold must be mapped from the Bark to the frequency scale. Additionally, the outer/middle ear transform must be reversed. Otherwise, the masking threshold will describe the conditions at the oval window rather than at the pinna.

The energy contained within one quarter-bark is distributed uniformly (in frequency) over the frequency bins which span it. The conversion algorithm is provided in greater detail in Appendix B.

Afterwards, the outer/middle ear transformation must be inverted. This can be done utilizing Eq. (3.2).

$$W_{\text{inv}}[i] = -W[i]\text{dB} \tag{4.15}$$

Clearly, W_{inv} will be singular at $f = 0$. Considering the auditory threshold of hearing, the masking threshold at that frequency can be set to any convenient value as this bin will be inaudible. The selected value will be $s_{w_0}^{(L)}$ causing the $f = 0$ bin to be unattenuated by the psychoacoustic filter.

4.7 Chapter Summary

This chapter focussed on the design of a perceptual post-filter to remove audible artefacts. Firstly, the phenomenon of musical noise were discussed in further detail. It was then shown how the principles of masking could be utilized to attenuate musical noise with minimal distortion. Finally, a detailed description of the perceptual post-filter and the resulting enhanced signal subspace algorithm were given.

Chapter 5

Experimental Results

This chapter will present an analysis of the performance of the Enhanced Signal Subspace (ESS) method. The implementation of the enhancement algorithm will be described. This will be followed by the results of objective and subjective testing. Finally, a qualitative study will be presented to better explain the characteristics of the proposed algorithm.

5.1 Algorithm Implementation

The implementation of the ESS algorithm requires the consideration of several issues. The calculation of the correlation matrix requires the selection of a estimator. The size and type of the analysis/synthesis windows must be considered. A linear estimator must also be chosen for the signal subspace filter.

5.1.1 Correlation Matrix Estimation

Correlation matrices were calculated using an estimator. It has the form

$$r_{u_{i,j}} = \frac{1}{N} \sum_{l=0}^{N-j-i-1} u_n u_{n+j-i} \tag{5.1}$$

where $\boldsymbol{u}$ and N denote the correlation data window contents and length, respectively. This estimator is biased as shown by

$$E\{r_{u_{i,j}}\} = \frac{N-|i-j|}{N} \gamma_{u_{i-j}} \tag{5.2}$$

and a variance of

$$\mathrm{VAR}\{r_{u_{i,j}}\} = \frac{1}{N} \sum_{n=-\infty}^{\infty} |\gamma_n|^2 + \bar{\gamma}_{u_{n-(i-j)}} \gamma_{u_{n+(i-j)}} \tag{5.3}$$

where $\gamma_{u_m} = E\{\bar{u}_i u_{i+m}\}$ and $\bar{\cdot}$ denotes the conjugation operation.

The estimator r_{u_m} is consistent. As the size of the correlation window increases, r_{u_m} will converge to the true autocorrelation value γ_{u_m}.

The lags of the signal correlation matrices are calculated using a rectangular window of 350 samples. The length of the window utilized in the correlation matrix estimation has a discernible effect on output speech. Longer windows will tend to reduce the variance of the higher order lags. Consequently, the estimated size of the signal subspace will remain relatively constant, diminishing the number of artefacts introduced from dimensionality changes. However, if an excessively large window is used to estimate the correlation matrix, the transients in the enhanced speech will be dulled.

The noise covariance matrix is updated during speech pauses. As the noise signal is assumed to have slowly varying statistics, this method is sufficient to obtain accurate estimates.

5.1.2 Selection of Analysis and Synthesis Windows

The signal subspace filter has been implemented with a frame size of 64 taps and 50% overlap. A rectangular analysis window is applied to the data prior to signal subspace filtering. The psychoacoustic filter operates on 350 sample frames. After application of the post-filter, a sine-squared synthesis window is utilized for reconstruction. This choice of windows produces maximum smoothing during reconstruction at the cost of spectral leakage during analysis.

5.1.3 Choice of Linear Estimator

Several linear estimators were proposed in Section 2.7 to approximate the uncorrupted signal. To determine the best estimator, informal listening tests were carried out. They were based on speech being corrupted by white noise with a Signal to Noise Ratio (SNR) of 10 dB. The following results were obtained:

- Least-Square Estimator: The output speech produced by the least-square estimator was found to contain a great deal of residual noise and musical tones. This could be attributed to the linear estimator not filtering the signal subspace.

- Minimum Variance Estimator: With $\gamma = 1$, the Minimum Variance (MV) estimator was found to reduce noise with the addition of a small amount of musical noise and a little distortion. Adjusting γ within the range of [0.5,1.5] did not produce audible benefits. It was also learned that the generalized Wiener filter aggressively attenuated noise but muffled output speech.

- Time Domain Constraint Estimator: The Time Domain Constraint (TDC) estimator is similar in form to the MV filter. Considering performance, the TDC estimator produced less musical noise than the MV filter without additional distortion.

The time domain constraint estimator was found to be the best choice. As discussed above, informal listening tests showed it to produce superior outputs.

5.2 Description of Test Data

To perform efficacy tests of the ESS algorithm, testing data was required. The recordings were obtained from a large speech database and resampled for further use. Noise was then added to simulate noisy environments.

5.2.1 Recording of Speech Files

The audio files utilized for testing were taken from the TSP speech database. This database is comprised of more than 24 speakers uttering over 1440 sentences. The recordings were made in an acoustic anechoic chamber with a Sony ECM-909A stereo microphone placed 15 cm from the speaker. The speech signals were stored on a Sony TCD-D3 DAT recorder at a sampling rate of 48 kHz and resampled to 8 kHz. To simulate a telephone channel, the narrowband speech was bandpass filtered between 200 and 3400 Hz. The channels were averaged to reduce noise.

The sentences were taken from the Harvard Sentence lists [72]. The Harvard sentences are a set of 100 utterances which are meaningful, syntactically varied, whose phonetic balance is similar to that of English language. Accordingly, performance measures based on these sentences will reflect the system in normal use.

5.2.2 Addition of Noise to Test Data

Noise was added to speech files to attain specific SNR. All noise samples were drawn from the NOISEX-92 database [73]. SNR was defined as the ratio of the active speech level to the root mean-square (RMS) level of the noise. This definition was taken from the ITU standard for Methods for Objective and Subjective Assessment of Quality (ITU-T P.830) [74].

$$\mathrm{SNR_{dB}} = 20\log_{10}\left\{\frac{\text{Active speech level}}{\text{RMS level of the noise}}\right\} \tag{5.4}$$

Such a definition ensures that speech pauses do not affect the calculation of the SNR. The active speech level corresponds to the following:

Definition 4 (Active Speech Level) *The active speech level is calculated by integrating a quantity proportional to the instantaneous power of the aggregate of time during which the speech in question is present, and then expressing the quotient, proportional to total energy divided by active time, in decibels relative to the appropriate reference [75].*

The active speech level will be denoted as

$$A(l_0) = 10\log_{10}\left\{\frac{\sigma_x^2 G_v^2}{a(l_0)}\right\} - R \tag{5.5}$$

where l_0 denotes the speech activity threshold and $a(l_0)$ represents the percentage of the speech signal greater than l_0. R and G_v are an offset and a scaling factor respectively. The log domain representation of the speech activity threshold is denoted as $C(l_0)$. It has the form

$$C(l_0) = 10\log_{10}\left\{l_0 G_v^2\right\} - R \tag{5.6}$$

It has been shown that a threshold of 15 dB below the RMS level of the signal is sufficient to separate active speech from silence [76]. Thus, the speech activity threshold must be chosen to satisfy

$$A(l_0) - C(l_0) = 10\log_{10}\left\{\frac{\sigma_x^2 G_v^2}{a(l_0)}\right\} - 10\log_{10}\left\{l_0 G_v^2\right\} = 15.9. \tag{5.7}$$

The additional 0.9 dB compensates for the difference between the mean absolute value and the root mean-square value of a sinusoid [77].

5.3 Qualitative Analysis of Enhanced Signal Subspace Method

A series of listening tests were performed to qualitatively describe the properties of the ESS algorithm. These trials examined the effect of the perceptual post-filter and the performance of the proposed algorithm under various noisy conditions.

5.3.1 Evaluation of the Efficacy of the Perceptual Post-filter

The effectiveness of the perceptual post-filter was determined by comparing the outputs of the speech enhancement algorithm with or without the presence of the perceptual post-filter.

Informal listening tests determined that without the perceptual post-filter the speech contained noticeable musical noise. When the perceptual post-filter was applied, these artefacts were largely attenuated, although, slight distortion was noticeable in the enhanced speech. These observations were made for speech corrupted by white noise with 10 dB SNR.

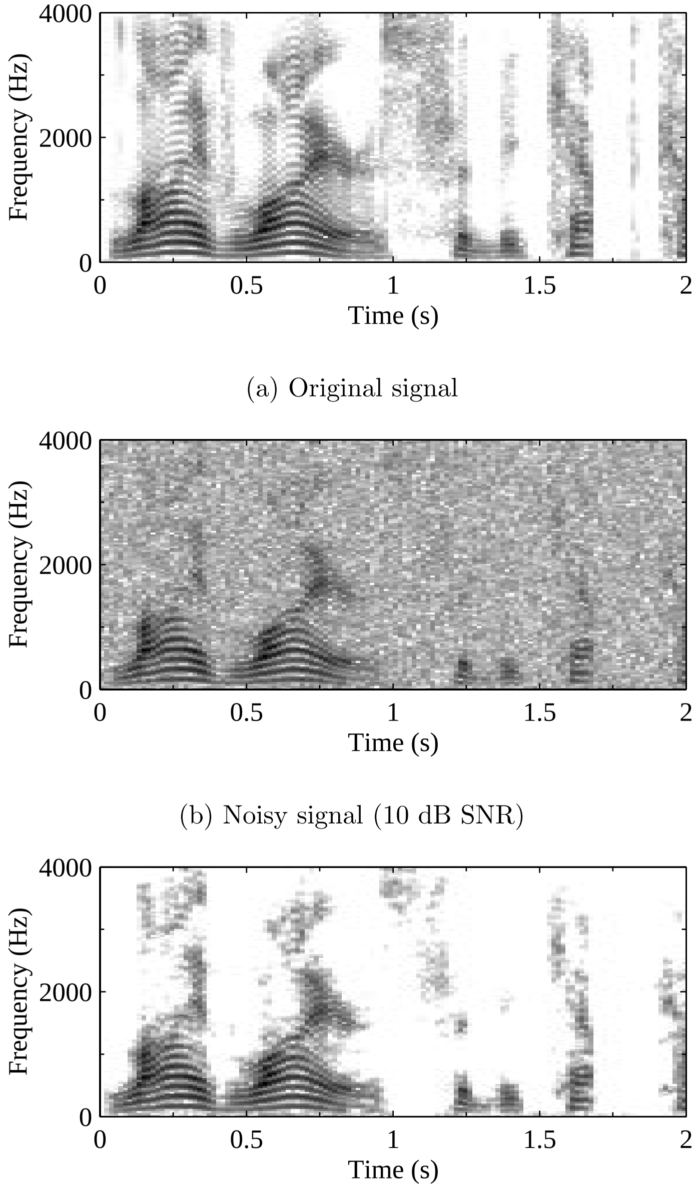
A spectrogram of the sentence “Live wires should be kept covered” is depicted in Fig. 5.1. It shows a speech file before and after enhancement with the proposed algorithm, as well as its clean state. The speech file has been perturbed by white noise and has an SNR of 10 dB.

It is evident from the denoised signal that the noise has been mostly removed. Unlike many enhancement algorithms which tend to muffle speech, this method retained the high frequency components. This ensured that the enhanced signal possessed naturalness.

It was also noted that voiced speech was better handled than unvoiced speech. This can be attributed to the innate suitability of voiced speech for low-rank representations. In contrast, unvoiced signals require near-full-rank models to be modelled accurately. Thus, larger noise subspaces can be removed from voiced speech to produce better enhancement results.

Further experimentation was performed by applying the perceptual post-filter to clean and noisy[1] speech files. It was found that the resultant files were not perceptibly differ-

[1] White noise was added to produce a 10 dB SNR.



(a) Original signal

(b) Noisy signal (10 dB SNR)

(c) Estimated signal

Fig. 5.1 Comparison of speech spectrograms

ent from their unfiltered state. This confirmed that the psychoacoustic filter affects only artefacts and does not audibly attenuate noise or speech.

5.3.2 Performance in White Noise

The performance of the system was examined under several different operating conditions. The SNR was assigned values of 6 dB, 10 dB, and 15 dB with additive white noise.

Table 5.1 Performance of enhancement scheme in white noise

SNR (dB)	Description
6	-Loss of naturalness -High intelligibility
10	-Voiced speech fully recovered -Unvoiced speech slightly distorted
15	-All speech fully recovered

When the high SNR signals of 15 dB were processed, the speech signal was recovered without distortion or artefacts. With little noise, the signal and noise subspaces could be easily separated. For an SNR of 10 dB, it was observed that only the voiced speech was recovered without error. For unvoiced speech, the weakest sounds suffered the most distortion. This was explained by the lack of suitability of unvoiced speech for signal subspace decomposition (see Section 2.2.3). Overall, the fidelity of the recovered speech was still quite high. A loss of naturalness was detected with the denoised 6 dB SNR sentences. Intelligibility, however, remained quite high for all test files. A summary of the testing results can be found in Table 5.1.

5.3.3 Performance in Coloured Noise

Coloured noise was then presented to the enhancement scheme to determine its flexibility with different noisy environments. Car noise, pink noise and destroyer engine noise were all used in testing. The signals were prewhitened for use with the signal subspace filter (see Section 2.8).

It was found that full speech recovery was possible with car noise. In contrast, the enhanced signal arising from pink noise displayed significant musical noise. Pink noise is defined as having an equal amount of energy per octave of bandwidth. Finally, the environmental noise of a destroyer engine room was employed. It was found that a significant amount of residual noise remained.

It was determined experimentally that these problems were produced by poor noise estimation. When the true noise correlation matrices were utilized, speech quality improved.

Table 5.2 Performance of enhancement scheme in coloured noise

Noise Type (dB)	Description
Pink Noise	-Low rumbling noise present -Slight distortion
Car Noise	-Noise fully removed
Destroyer Engine Noise	-Residual noise still prevalent

5.4 Analysis of Order Estimators

A reasonable estimate of the signal subspace dimensionality is critical to the operation of the signal subspace filter. Consequently, the order estimator must be robust in noise and able to provide values for full rank systems. To choose among the algorithms proposed in Section 2.7, two experiments were carried out:

- Processing of a fixed order signal in noise: The order estimators are applied to a signal of fixed order in varying amounts of noise. This trial assessed the robustness of the measure to noise.

- Processing of speech signals in noise: The respective estimators are applied to the set of speech files with very high SNR. The estimator should be able to be decompose speech into signal and noise subspaces without disturbing quality. The order estimates for each frame are compared to the subjective quality of the resultant files.

5.4.1 Processing of a fixed order signal in noise

To assess the reliability of the order estimator, the prospective algorithms were applied to a signal of known order in additive white noise. The power of the noise signal was increased to examine the performance of the estimator in noise.

The chosen test signal was comprised of five sinusoids normalized to have an average power of 1. The signal had a dimensionality of ten as an individual sinusoid can be modelled using a second-order difference equation. The variance of the noise was varied to measure the reliability of the method. Fig. 5.2 depicts the estimates provided by three order estimation methods.

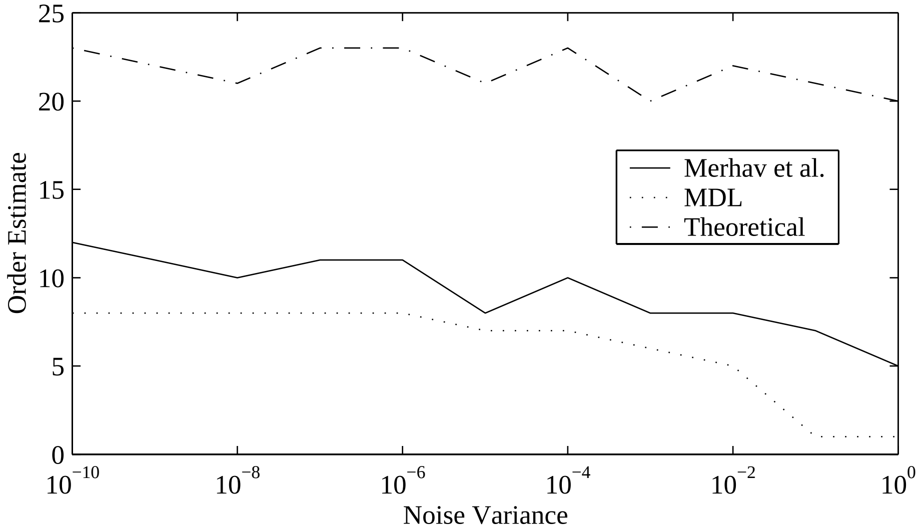


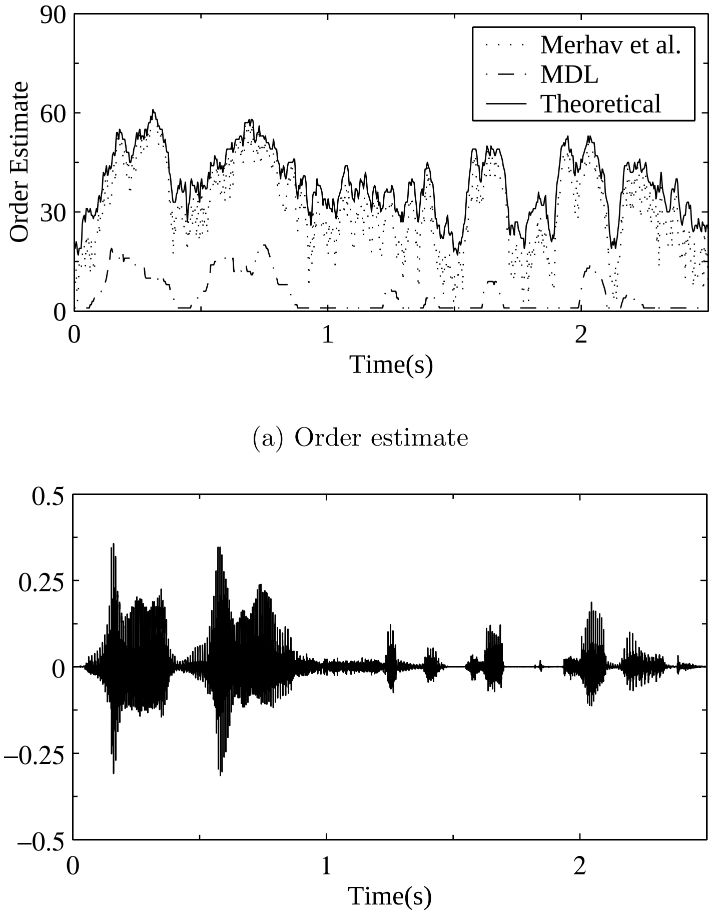
Fig. 5.2 Order estimate of sinusoids in noise

Experimental results showed that the theoretical estimator tended to greatly overestimate signal order. This large error could be attributed to ringing in the eigenvalue estimates. The theoretical estimator is known to be extremely sensitive to this phenomenon. Using smoother windows was shown to reduce oscillations and produce better results. Improvements could have been also achieved by using a threshold slightly greater than zero to determine rank.

The MDL and Merhav *et al.* estimators were both observed to provide accurate estimates for a wide range of noise variances.

5.4.2 Processing of speech signals in noise

The estimators were applied to a speech signal with a high SNR. This experiment determined if the estimator could decompose full rank uncorrupted signals into signal and noise subspaces. In terms of enhanced speech quality, it is better to overestimate signal order than underestimate it. The quality of the resultant enhanced speech will be used as a criteria to help select an estimation method.



(a) Order estimate

(b) Time signal

Fig. 5.3 Comparison of order estimation algorithms

Fig. 5.3 depicts the order estimators for a typical speech signal corrupted with white noise possessing an SNR of 100 dB. The MDL algorithm was shown to be very aggressive, eliminating over 75% of the vector space. The theoretical estimator produced the highest order estimates, while the Merhav *et al.* algorithm produced values which were slightly smaller.

From informal listening tests, it was concluded that the theoretical estimator and the Merhav *et al.* algorithm sounded identical. They both performed well at estimating speech order. The MDL estimator tended to produce distorted speech by eliminating too much speech.

It was noted that the order estimators behaved similarly over the same test sentences. The dimensionality estimates dropped during the speech pauses. Surprisingly, unvoiced speech tended to produce a low order estimate. From the randomness of unvoiced speech, it would be expected that this class of phoneme should require a high order model. However, poor estimates of weaker eigenvalues made it impossible to use a higher dimensionality model.

5.4.3 Order Estimator Choice

The theoretical estimator was chosen for the ESS algorithm. While the algorithm was shown to be biased for estimating the order of signals with fixed order, it remained consistent over a wide range of noise variances. Furthermore, the order estimator was found to be effective for actual signals.

5.5 AB Testing

AB testing was performed to measure the quality of the proposed enhancement algorithm. Subjective testing presents the clearest indication of quality for a speech enhancement algorithm. The ESS output was compared to several enhancement schemes including the noise suppression block of the Enhanced Variable Rate Coder (EVRC).

5.5.1 Evaluation Protocol

AB testing was by interviewing ten subjects. Listeners were presented identical files having been processed in two different manners. They were asked to choose their preference, if they had one.

The test data was compromised of two sentences separated by a one second pause. Each sentence was spoken by the same speaker. The order of the two processing methods were randomized to prevent listener bias. An equal number of male and female speakers were presented to prevent any gender skew. The volume was adjusted by the listeners to their preference.

The AB testing was divided into four trials. Each trial compared the ESS algorithm to a different method. A total of eight speech files were used for each test.

5.5.2 Results of AB Testing

The speech files under analysis were corrupted by white noise with an SNR of 10 dB. Table 5.3 and Table 5.4 show the results of AB testing.

Table 5.3 Percentage of listener preference

	Listener Preference		
Method	ESS	Undecided	Other Method
Trial 1: ESS / Noisy Signal	70%	0%	30%
Trial 2: ESS / Ephraim and Van Trees	80%	10%	10%
Trial 3: ESS / EVRC	60%	0%	40%
Trial 4: ESS / Spectral Subtraction	30%	10%	60%

Table 5.4 Number of votes cast for each algorithm

	Number of Votes		
Method	ESS	Undecided	Other Method
Trial 1: ESS / Noisy Signal	40	12	28
Trial 2: ESS / Ephraim and Van Trees	60	7	13
Trial 3: ESS / EVRC	97	29	34
Trial 4: ESS / Spectral Subtraction	21	21	38

Trial 1 *Noisy speech compared to ESS*: The subjects were presented noisy and ESS enhanced speech files. It was shown that a majority of subjects preferred enhanced speech to the original noisy files. This test proved that the enhancement algorithm is beneficial and that it improves speech quality.

Trial 2 *Ephraim & Van Trees compared to ESS*: The listeners were asked to compare speech files processed with the Ephraim & Van Trees algorithm and those which passed through the ESS algorithm. This experiment assessed the efficacy of the psychoacoustic filter. The results showed that the ESS method was superior to Ephraim and Van Trees algorithm. It can be concluded that the musical noise reduction of ESS is effective.

Trial 3 *EVRC Noise Suppression Block compared to ESS*: The output of the EVRC noise suppression block was compared to that of ESS. The results show that the proposed algorithm was slightly preferred. It may be concluded that listeners preferred slight distortion to significantly increased residual noise.

A description of the EVRC noise suppression block can be found in Appendix C.

Trial 4 *Spectral Subtraction compared to ESS*: The spectral subtraction method was tested with the ESS algorithm for analysis. Testing showed that spectral subtraction was more appropriate for white noise.

It was surprising that spectral subtraction outperformed ESS. Thus, additional listening tests were performed using coloured noise. Informal results showed that ESS outperformed spectral subtraction in this instance.

5.6 Mean Opinion Score Testing

Mean Opinion Score (MOS) quantify the performance of speech algorithms by assigning a value from 1 to 5. The MOS system is defined using "anchors" described in ITU-R BS.1116-1 [78] and reproduced in Table 5.5.

Table 5.5 Five-grade impairment scale

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very Annoying	1.0

The testing conditions necessary for proper MOS testing are quite stringent. Hence, an objective measure was utilized to generate consistent results. Objective testing was carried out using the Perceptual Evaluation of Speech Quality (PESQ) algorithm outlined in ITU standard (ITU-T P.862) [79]. The algorithm outputs a value similar to MOS.

5.6.1 Description of PESQ Algorithm

The PESQ model was designed to evaluate 3.1 kHz (narrow-band) handset telephony speech codecs. The PESQ algorithm requires both the degraded signal and a reference signal for comparison. An overview of the PESQ algorithm has been depicted in Fig. 5.4.

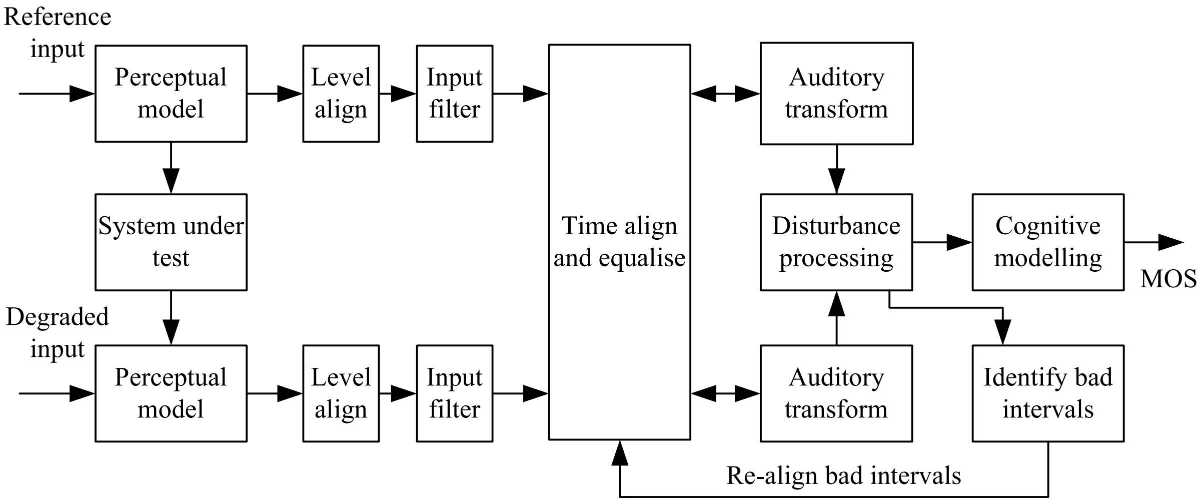


Fig. 5.4 Overview of the PESQ algorithm

The quality measure takes into account filtering and variable delay, allowing application across a wider range of network conditions. While PESQ has not been validated for use with speech enhancement, its design indicates that it would lend itself well to this purpose. The PESQ algorithm consists of several distinct functional stages. They can be classified as:

- Signal pre-processing: The signals under analysis are level aligned to a fixed listening volume. They are also time aligned to allow comparisons between corresponding parts of the original and degraded signals.

- Application of Perceptual Model: The auditory transformation maps the signals into a perceived loudness scale. The magnitude response of the spectra is grouped into 42 groupings spaced in perceptual frequency. Short-term energy fluctuations are clipped and smoothed over time. The groupings are mapped to a Sone loudness scale using Zwicker's transform law.

- Calculation of Distortion Measures: The absolute difference between the loudness of the degraded and reference signals are utilized to give a measure of perceived distortion. Two measures are produced from this stage, an asymmetric and a symmetric measure.

  The asymmetry factor is utilized to model the perceptual difference between additions and deletions. The ear is much more sensitive to additions than deletions. It is more difficult to introduce a time-frequency component that integrates with the input signal. Accordingly, the output will be decomposed into distinct auditory objects. Deletion does not produce such effects.

- Mapping to MOS scale: The asymmetric and symmetric distortion measures are mapped to a MOS-like scale using a linear transformation.

5.6.2 Results of PESQ simulations

The PESQ algorithm was applied to enhancement of ten sentences from the Harvard sentence list. An equal number of male and female speakers were utilized. The test data set was corrupted with white noise to obtain an SNR of 10 dB. It was determined that the ESS method outperformed all others.

Fig. 5.5 shows the results averaged by talker gender. Noisy speech was shown to be the most disturbing while spectral subtraction and EVRC had MOS lying between the two extremities. ESS had the highest ratings.

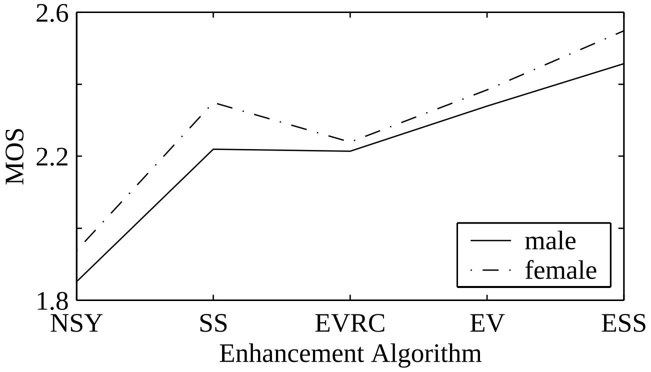


Fig. 5.5 Overview of the PESQ algorithm (NSY - Noisy Speech / SS - Spectral Subtraction / EVRC - Enhanced Variable Rate Coder / EV - Ephraim and Van Trees / ESS - Enhanced Signal Subspace method)

5.7 Chapter Summary

This chapter examined the performance and design of the Enhanced Signal Subspace method. The implementation of the ESS algorithm was first carefully described. Afterwards, a qualitative analysis of the algorithm was presented. Finally, the outcome of subjective and objective trials was discussed.

Chapter 6

Conclusion

This thesis has focussed on the design and implementation of a speech enhancement algorithm based on signal subspace techniques. The proposed algorithm is denoted as the Enhanced Signal Subspace (ESS) method. It improves over the traditional subspace framework by employing a perceptual post-filter. Artefacts are eliminated from the enhanced signal by exploiting the properties of the human ear. The resulting speech has been smoothed such that disturbances are removed without introducing perceptible distortion.

This chapter will provide a summary of the thesis work. Additionally, directions for future work will be discussed.

6.1 Summary of Work

The Enhanced Signal Subspace algorithm was created to remove noise from speech signals without introducing musical noise or distortion. The speech enhancement system was based on the signal subspace techniques described by Ephraim and Van Trees in [16]. Listening tests showed that their algorithm alone produced disturbing musical noise. These disturbances stemmed from sudden changes in signal subspace dimensionality and poor estimation of signal parameters.

Musical noise was shown (in Section 4.1) to be characterized by tones switching on and off at different frequencies. In part, it could be noticeably reduced by utilizing large data windows to calculate correlation matrices. To completely suppress the disturbances, a perceptual post-filter was employed. It shaped the residual noise to reduce the detectability of musical tones while leaving the underlying speech signals undisturbed. Intuitively, this could be accomplished by filtering the estimated noise residual at the output of the signal subspace system to be below the masking threshold. However, it was empirically determined that shaping the noise residual to resemble the clean speech spectrum was sufficient. Attempting to place the signal subspace noise below the masking threshold was found to produce unpleasant outputs. The problem originated from unreliable estimates of the residual noise. The estimator of the power spectral density of a nonstationary signal is non trivial.

The strength of the perceptual post-filter lay in its utilization of the concept of masking to minimize attenuation. By filtering only what is audible, artefacts are reduced. Furthermore, the perceptual filter produced spectral and temporal averaging to smooth the noise residual. The process of spectral averaging increases the width of tones within the noise residual. Temporal averaging, on the other hand, smoothes magnitude changes of the noise residual over several frames. Both processes effectively reduce the saliency of musical noise.

An extensive set of experiments were performed to examine the properties of the ESS algorithm. It was determined that the algorithm consistently improved the quality of signals at even low SNR. PESQ, a quality measure proposed by the ITU [79], showed that ESS compared favourably with contemporary algorithms such as spectral subtraction and the noise suppression block of EVRC.

To further improve enhanced speech quality, the properties of the signal subspace framework were carefully examined. An emphasis was placed on selecting an accurate order estimator and an appropriate linear estimator. From this study, it was shown that the theoretical order estimator and Time Domain Constraint estimator were superior.

It was experimentally verified that voiced speech was well suited for signal subspace enhancement. These utterances were well represented by low-rank models. Accordingly, large noise subspaces were able to be discarded without producing speech distortion. Contrarily, unvoiced speech was found to require near full-rank models. Unfortunately, inaccurate estimates of weaker eigenvalues prevented the use of higher order models. This was found to be one of the greatest limitations of signal subspace based algorithms.

Many signal enhancement methods suffer from difficulties with parameter estimation. Spectral subtraction, for example, has difficulty restoring weak spectral bins in strong noise. Known estimators are insufficient to provide accurate results. Signal subspace methods have an advantage over alternative enhancement methods. By performing principal component analysis, the signal energy is concentrated into a minimal number of transform coefficients.

These values will fare better in the estimation process than other transform coefficients.

It was found in the implementation of the perceptual post-filter that varying attenuation allowed a trade-off between musical noise and distortion. The distortion associated with the ESS algorithm was similar to low-pass filtering. As the AB testing showed, listeners prefer distortion to musical noise. Thus, the artefacts were aggressively removed. Unfortunately, if attenuation was too great, speech quality began to suffer. Intelligibility became impaired as fricatives became difficult to distinguish.

The ESS algorithm performed well removing coloured noise. However, the technique had some difficulty with the removal of noise with a large bandwidth. An audible noise residual was detected in the enhanced speech. This difficulty was attributable to the noise estimation and prewhitening blocks. It was determined that using the actual noise covariance matrix allowed complete removal of the noise.

6.2 Future Work

This section will detail future directions for work on the Enhanced Signal Subspace algorithm. The main issues to be considered include reduction of computational complexity, improved coloured noise handling and addition of a spectral floor. Also, attention will be placed on applying the proposed method to a more general class of signals.

6.2.1 Reduction of Computational Complexity

The ESS algorithm is quite complex. Signal subspace decomposition and the PEAQ masking model each have a computational complexity of $\mathrm{O}(n^3)$. These heavy requirements limit the practical applications of the ESS algorithm. Fortunately, several options exist to simplify the speech enhancement algorithm. They include using signal independent transforms and less complex masking models.

The Karhunen-Loève Transform (KLT) provides an optimal representation of a speech signal in terms of energy compaction. However, several signal independent transforms have shown similar decorrelative properties for speech signals. The Discrete Cosine Transform (DCT), for example, is known to asymptotically approach the performance of the KLT as the frame size becomes large. This property was extensively studied by Sánchez *et al.* in [80]. They obtained expressions for the decorrelating behaviour of the DCT with respect to an arbitrary stationary process. In [23], Huang and Zhao showed that the DCT was practical for subspace decomposition.

To further simplify the implementation of the ESS algorithm, a less exact masking model could have been employed. Considering estimator uncertainty for the clean speech spectrum, a high precision model is not required. Less exacting models may still be employed to produce high quality output. Viable alternatives include the Johnston model [65] and the MPEG-1 psychoacoustic model I [81].

6.2.2 Experimentation with Other Coloured Noise Handling Methods

To process coloured noise, a prewhitening block was applied to the input signal. As described in Section 2.8, this operation is performed using the inverse root of the noise correlation matrix. Unfortunately, this prewhitening method was shown to not produce optimal enhancement filters by Mittal and Phamdo in [26]. The noise correlation matrix also suffers from ill-conditioning in the presence of narrowband noise. Thus, alternative methods for coloured noise handling should be explored.

One prospective method was described by Rezayee and Gazor in [24]. The speech eigenvalues are calculated from an estimate of the speech correlation matrix. Unfortunately, the resultant eigenvectors are not be able to diagonalize the noise correlation matrix. Thus, the noise eigenvalues are approximated using the diagonal entries of the similar matrix. It is assumed that the off-diagonal components contain negligible amounts of energy. It has been shown that this method provides good results.

6.2.3 Introduction of a Spectral Floor to Mask Distortion

Enhancement algorithms such as the noise suppression block in the EVRC algorithm are effective because they incorporate a noise floor. By leaving a noise residual, severe distortion is prevented and artefacts are masked. This technique could prove useful in the ESS algorithm. It may be incorporated by decreasing attenuation in the psychoacoustic filter or adding noise into the output of the signal subspace filter.

At present, the psychoacoustic filter attenuates until the noise contained in the corrupted signal lies below the masking threshold. However, it may be more advantageous to stop several dB above the audibility threshold. The noise spectrum will be shaped such that the peaks have been widened and interframe fluctuations will have been minimized.

However, a small amount of residual noise will have been introduced to mask distortions

Alternatively, noise could be reintroduced at the output of the signal subspace filter. This can be accomplished by filtering white noise with the normalized power spectral density of the original noise and adding it to the enhanced signal.

Preliminary listening tests have shown that both methods are promising. The noise residual in both cases masked distortion while producing speech that was pleasing to the listener.

6.2.4 Application to General Audio

The ESS algorithm has been designed to process speech. The enhancement method, however, is not limited to this group of signals. The requirements for the signal subspace filter and the perceptual post-filter are flexible enough that a much larger class of signals may be accommodated.

Signal subspace methods can be applied to any signal which permits a low-rank representation. The benefits of the enhancement technique increases as the signal dimensionality decreases. Acoustic instruments tend to possess spectra composed of discrete harmonics. Accordingly, they lend themselves to being represented with a reduced dimensionality linear model. Thus, many types of music, including classical, are well suited for the ESS algorithm.

The perceptual post-filter can be used on an even larger class of signals. The system simply calculates a masking threshold of the estimated clean signal. From this audibility threshold, filter coefficients are derived. This operation clearly does not necessitate any structure in the audio sample under analysis.

Appendix A

Derivation of Linear Estimators

This appendix will detail the derivation of the linear estimators described in Section 2.7. To obtain closed form expressions, the method of Lagrange multipliers for inequality constraints will be applied.

The first section will introduce the Lagrange multiplier method, while subsequent section will detail the derivation of the estimators.

A.1 Necessary and Sufficient Conditions for the Solution of an Inequality Constrained Optimization Problem

The necessary and sufficient conditions for the solution of an inequality constrained optimization problem will now be presented [82]. It will be assumed that the minimization problem has the form

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ \text{subject to: } &g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \dots, m \\ &h_i(\boldsymbol{x}) = 0, \quad i = 1, \dots, p \end{aligned} \tag{A.1}$$

where the functions $f, g_1, \dots, g_m, h_1, \dots, h_p$ are defined and differentiable over some open set $D \subset \mathbb{R}^N$.

The Lagrangian is defined as

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \mu_j h_i(\boldsymbol{x}). \tag{A.2}$$

A stationary point, $\bar{\boldsymbol{x}}$, satisfies Eq. (A.3). A minima for an optimization problem must be a stationary point.

$$\begin{gathered}\nabla_x L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}) = \nabla f(\bar{\boldsymbol{x}}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(\bar{\boldsymbol{x}}) + \sum_{j=1}^{p} \bar{\mu}_j \nabla h_i(\bar{\boldsymbol{x}}) = 0 \\ \bar{\lambda}_i g_i(\bar{\boldsymbol{x}}) = 0, \quad i = 1, \dots, m \\ \bar{\boldsymbol{\lambda}} > \mathbf{0}\end{gathered} \tag{A.3}$$

Let F denote the set of values which satisfy the optimization constraints. A feasible direction vector is defined as a vector such that there exists a δ where if $\boldsymbol{w} \in F$ then $(\boldsymbol{w} + \alpha \boldsymbol{z}) \in F$ for $0 \leq \alpha \leq \delta$. $Z(\bar{\boldsymbol{w}})$ is the set of all feasible directions vectors at $\boldsymbol{w}$.

If a stationary point, $\bar{\boldsymbol{x}}$, meets Eq. (A.4) for every $\boldsymbol{z} \neq \mathbf{0}$ where $\boldsymbol{z} \in Z(\boldsymbol{x}^*)$, then it is a strict local minima of the problem.

$$\boldsymbol{z}^T [\nabla^2 f(\bar{\boldsymbol{x}}) - \sum_{i=1}^{m} \bar{\lambda}_i \nabla^2 g_i(\bar{\boldsymbol{x}}) - \sum_{j=1}^{p} \bar{\mu}_j \nabla^2 h_i(\bar{\boldsymbol{x}})] \boldsymbol{z} > 0 \tag{A.4}$$

The solution of an optimization problem will be denoted as $\boldsymbol{x}^*$

A.2 Derivation of Time Domain Constraint Filter

The Time Domain Constraint (TDC) estimator results from the solution of the following optimization problem

$$\begin{gathered}\boldsymbol{H}^* = \arg \min_{\boldsymbol{H}} \epsilon_x^2 \\ \text{subject to: } \frac{1}{M} \epsilon_w^2 \leq \alpha \sigma_w^2.\end{gathered} \tag{A.5}$$

Recall:

$$\epsilon_x^2 = \operatorname{tr}\{(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{R}_x(\boldsymbol{H}-\boldsymbol{I})^H\} \tag{A.6}$$

$$\epsilon_w^2 = \sigma_w^2 \operatorname{tr}\{\boldsymbol{H}\boldsymbol{H}^H\}. \tag{A.7}$$

To obtain a closed form solution, the Lagrangian must be calculated. It is found to have the form

$$\begin{aligned} L(\boldsymbol{H},\lambda) &= \epsilon_x^2 + \lambda(\frac{1}{M}\epsilon_w^2 - \alpha\sigma_w^2) \\ &= \operatorname{tr}\{(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{R}_x(\boldsymbol{H}-\boldsymbol{I})^H\} + \lambda(\frac{\sigma_w^2}{M}\operatorname{tr}\{\boldsymbol{H}\boldsymbol{H}^H\} - \alpha\sigma_w^2). \end{aligned} \tag{A.8}$$

Calculating the gradient of the Lagrangian and the Hessian, the following is obtained

$$\nabla L(\boldsymbol{H},\lambda) = 2(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{R}_x + 2\alpha\sigma_w^2\boldsymbol{H} \tag{A.9}$$

$$\nabla^2 L(\boldsymbol{H},\lambda) = 2\boldsymbol{R}_x + 2\alpha\sigma_w^2\boldsymbol{I}. \tag{A.10}$$

As the Hessian is positive definite, any stationary point will be a strict local minima. Thus, the optimal filter will be

$$\boldsymbol{H}^* = \boldsymbol{R}_x(\boldsymbol{R}_x + \lambda\sigma_w^2\boldsymbol{I})^{-1}. \tag{A.11}$$

It remains to determine the value of the Lagrangian multiplier. If the constraint is assumed to be active (i.e. $\lambda \neq 0$), then

$$\epsilon_w^2 = \alpha M \sigma_w^2. \tag{A.12}$$

Substituting Eq. (2.42) in Eq. (A.12), one obtains

$$\alpha = \sigma_w^2 \operatorname{tr}\{\boldsymbol{R}_x^2(\boldsymbol{R}_x + \lambda\sigma_w^2\boldsymbol{I})^{-2}\} \tag{A.13}$$

A.3 Derivation of Minimum Variance Filter

The minimum variance estimator problem can be expressed in the following manner:

$$\boldsymbol{H}^* = \arg\min_{\boldsymbol{H}} \epsilon_x^2 + \epsilon_w^2. \tag{A.14}$$

Again, the Lagrange multipliers will be used to determine the optimal solution. Therefore, the Lagrangian and its gradient must be first evaluated

$$L(\boldsymbol{H},\lambda) = \operatorname{tr}\{(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{R}_x(\boldsymbol{H}-\boldsymbol{I})^H\} + \sigma_w^2 \operatorname{tr}\{\boldsymbol{H}\boldsymbol{H}^H\} \tag{A.15}$$

$$\nabla L(\boldsymbol{H},\lambda) = 2(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{R}_x + 2\sigma_w^2\boldsymbol{H}. \tag{A.16}$$

The Hessian is shown to be positive definite in Eq. (A.17). Thus, any stationary point will be a minima.

$$\nabla^2 L(\boldsymbol{H},\lambda) = 2\boldsymbol{R}_x + 2\sigma_w^2\boldsymbol{I}, \tag{A.17}$$

Setting the gradient of the Lagrangian to zero, the optimal solution H^* is found to be

$$\boldsymbol{H}^* = \boldsymbol{R}_x(\boldsymbol{R}_x + \sigma_w^2\boldsymbol{I})^{-1}. \tag{A.18}$$

Applying the similarity transformation $\boldsymbol{R}_x = \boldsymbol{Q}\boldsymbol{\Lambda}_x\boldsymbol{Q}^H$ Eq. (A.19) is obtained.

$$\begin{aligned}\boldsymbol{H}^* &= \boldsymbol{Q}\Lambda_x(\Lambda_x + \sigma_w^2\boldsymbol{I})^{-1}\boldsymbol{Q}^H \\ &= \boldsymbol{Q}_1\Lambda_{x_1}(\Lambda_{x_1} + \sigma_w^2\boldsymbol{I})^{-1}\boldsymbol{Q}_1^H\end{aligned} \tag{A.19}$$

A.4 Derivation of the Auditory Filter

The auditory filter minimizes signal distortion while constraining the spectrum of the noise residual to be beneath the masking threshold. Thus, it satisfies

$$\begin{aligned}\boldsymbol{H}^* &= \arg\min_{\boldsymbol{H}} \; E\{\|(\boldsymbol{H}\boldsymbol{F}^H - \boldsymbol{F}^H)\boldsymbol{x}\|_2^2\} \\ &\text{subject to: } E\{\|\boldsymbol{f}_i^H\boldsymbol{H}\boldsymbol{w}\|_2^2\} \le m_i^2 \\ \boldsymbol{H}^* &= \operatorname{diag}(h_1,\ldots,h_N)\end{aligned} \tag{A.20}$$

where $\boldsymbol{F}$ denotes the Fourier transform matrix, $\boldsymbol{m}$ represents the masking threshold and $\boldsymbol{R}_w$ represents the noise correlation matrix. Finally, $\boldsymbol{x}$ indicates the signal being filtered and N specifies the FFT size.

A.4.1 Simplification of Objective and Constraint Functions

The objective function can be reexpressed in the following manner

$$\begin{aligned} f(\boldsymbol{H}) &= E\{\mathrm{tr}\{(\boldsymbol{H}\boldsymbol{F}^H\boldsymbol{x} - \boldsymbol{F}^H\boldsymbol{x})(\boldsymbol{x}^H\boldsymbol{F}\boldsymbol{H}^H - \boldsymbol{x}^H\boldsymbol{F})\}\} \\ &= E\{\mathrm{tr}\{\boldsymbol{H}\boldsymbol{F}^H\boldsymbol{x}\boldsymbol{x}^H\boldsymbol{F}\boldsymbol{H}^H\} + \mathrm{tr}\{\boldsymbol{F}^H\boldsymbol{x}\boldsymbol{x}^H\boldsymbol{F}\} - 2\,\mathrm{tr}\{\boldsymbol{H}\boldsymbol{F}^H\boldsymbol{x}\boldsymbol{x}^H\boldsymbol{F}\}\} \\ &= \mathrm{tr}\{\boldsymbol{H}\boldsymbol{F}^H\boldsymbol{R}_x\boldsymbol{F}\boldsymbol{H}^H\} + \mathrm{tr}\{\boldsymbol{F}^H\boldsymbol{R}_x\boldsymbol{F}\} - 2\,\mathrm{tr}\{\boldsymbol{H}\boldsymbol{F}^H\boldsymbol{R}_x\boldsymbol{F}\} \\ &= \mathrm{tr}\{\boldsymbol{H}\boldsymbol{S}_x\boldsymbol{H}^H\} + \mathrm{tr}\{\boldsymbol{S}_x\} - 2\,\mathrm{tr}\{\boldsymbol{H}\boldsymbol{S}_x\} \end{aligned} \tag{A.21}$$

where $S_w = \boldsymbol{F}^H\boldsymbol{R}_w\boldsymbol{F}$.

The constraints will now be proven to be independent. Consider the following simplification

$$\begin{aligned} E\{|\boldsymbol{f}_i^H\boldsymbol{r}_w|^2\} &= E\{\boldsymbol{f}_i^H\boldsymbol{F}\boldsymbol{H}\boldsymbol{F}^H\boldsymbol{w}\boldsymbol{w}^H\boldsymbol{F}\boldsymbol{H}^H\boldsymbol{F}^H\boldsymbol{f}_i\} \\ &= \boldsymbol{f}_i^H\boldsymbol{F}\boldsymbol{H}\boldsymbol{F}^H\boldsymbol{R}_w\boldsymbol{F}\boldsymbol{H}^H\boldsymbol{F}^H\boldsymbol{f}_i \\ &= \boldsymbol{f}_i^H\boldsymbol{F}\boldsymbol{H}\boldsymbol{S}_w\boldsymbol{H}^H\boldsymbol{F}^H\boldsymbol{f}_i \\ &= \boldsymbol{e}_i^H\boldsymbol{H}\boldsymbol{S}_w\boldsymbol{H}^H\boldsymbol{e}_i \\ &= s_{w_{ii}}h_i^2 \end{aligned} \tag{A.22}$$

where $S_w = \boldsymbol{F}^H\boldsymbol{R}_w\boldsymbol{F}$ and $\boldsymbol{e}_i$ denotes the elementary vector. Therefore, the constraint functions can be rewritten as

$$h_i \le \frac{m_i}{(s_{w_{ii}})^{\frac{1}{2}}} \tag{A.23}$$

A.4.2 Solution of Simplified Problem

Due to the independence of the constraints, the optimization problem can be restated as N inequality constrained optimization problems.

$$\begin{aligned} h_i^* = \arg\min_{h_i}\ & s_{x_{ii}}h_i^2 - 2\sum_i h_i s_{x_{ii}} + \sum_i s_{x_{ii}} \\ \text{subject to: } & h_i \le \frac{m_i}{(s_{w_{ii}})^{\frac{1}{2}}} \end{aligned} \tag{A.24}$$

The objective function provided in Eq. (A.24) is monotonically decreasing in the interval $[0, 1]$. As such, the largest feasible point is the optimal solution. Thus,

$$h_i^* = \begin{cases} 1 & m_i \geq (s_{w_{ii}})^{\frac{1}{2}}, \\ \dfrac{m_i}{(s_{w_{ii}})^{\frac{1}{2}}} & m_i < (s_{w_{ii}})^{\frac{1}{2}} \end{cases} . \tag{A.25}$$

Appendix B

Power Grouping

This section will provide the power grouping algorithm utilized in the PEAQ standard [58] to integrate power in quarter-bark bands. The inverse operation will also be presented.

In the ensuing, $F_l[i]$ and $F_u[i]$ will denote the lower and upper frequency boundaries of the i^{th} quarter-bark. F_{res} will denote the frequency resolution of the FFT utilized. Finally, $F_{\text{sp}}[k]$ will represent the power in the k^{th} frequency bin while $P_e[i]$ will correspond to the power in the i^{th} quarter-bark.

B.1 Quarter-Bark Power Grouping

This algorithm integrates the power contained a quarter-bark.

```
for $i = 0$; $i < Z$; $i$++ do
  for $k = 0$; $i < N$; $k$++ do

    if $(k - 0.5)F_{\text{res}} \geq F_l[i]$ and $(k + 0.5)F_{\text{res}} \geq F_u[i]$ then
      $P_e[i]$ += $F_{sp}[k]$

    else if $(k - 0.5)F_{\text{res}} < F_l[i]$ and $(k + 0.5)F_{\text{res}} > F_u[i]$ then
      $P_e[i]$ += $F_{sp}[k]\dfrac{(F_u[i] - F_l[i])}{F_{\text{res}}}$

    else if $(k - 0.5)F_{\text{res}} < F_l[i]$ and $(k + 0.5)F_{\text{res}} > F_l[i]$ then
      $P_e[i]$ += $F_{sp}[k]\dfrac{((k + 0.5)F_{\text{res}} - F_l[i])}{F_{\text{res}}}$
```

```
        else if $(k - 0.5)F_{\text{res}} < F_u[i]$ and $(k + 0.5)F_{\text{res}} > F_u[i]$ then
            $P_e[i]$ += $F_{sp}[k]\dfrac{(F_u[i] - (k - 0.5)F_{\text{res}})}{F_{\text{res}}}$

        end if

    end for
end for
```

B.2 Frequency Bin Power Grouping

This algorithm estimates the power contained within a frequency bin. It is assumed that the power contained within each quarter-bark is distributed uniformly in frequency.

```
for $i = 0$; $i < Z$; $i$++ do
    for $k = 0$; $i < N$; $k$++ do

        if $(k - 0.5)F_{\text{res}} \leq F_l[i]$ and $(k + 0.5)F_{\text{res}} \geq F_u[i]$ then
            $M_f[k]$ += $M_b[k]$

        else if $(k - 0.5)F_{\text{res}} > F_l[i]$ and $(k + 0.5)F_{\text{res}} < F_u[i]$ then
            $M_f[k]$ += $M_b[k]\dfrac{F_{\text{res}}}{(F_u[i] - F_l[i])}$

        else if $(k - 0.5)F_{\text{res}} < F_l[i]$ and $(k + 0.5)F_{\text{res}} > F_l[i]$ then
            $M_f[k]$ += $M_b[k]\dfrac{((k + 0.5)F_{\text{res}} - F_l[i])}{(F_u[i] - F_l[i])}$

        else if $(k - 0.5)F_{\text{res}} < F_u[i]$ and $(k + 0.5)F_{\text{res}} > F_u[i]$ then
            $P_e$ += $M_b[k]\dfrac{(F_u[i] - (k - 0.5)F_{\text{res}})}{(F_u[i] - F_l[i])}$

        end if

    end for
end for
```

Appendix C

Description of EVRC Noise Suppression Block

The noise suppression block of the EVRC algorithm was among the methods compared to the ESS technique. The EVRC system was developed by Motorola as part of the TIA/EIA standard IS-127 [83] for CDMA-based telephone systems.

C.1 Description of EVRC Algorithm

The EVRC algorithm operates with a frame length of 128 samples. These frames are compromised of 80 samples from the current frame, 24 samples from the previous frame and 24 zeros. The data is multiplied by a smoothed trapezoidal window, pre-emphasized, and transformed to the frequency domain with a 128-point FFT. The spectral coefficients are grouped into 16 *channels* to model the critical bands of the human ear (see Section 3.3.3). The SNR in each channel is calculated using the channel energy estimate and background energy statistics. The ratios are then converted to the log-domain and quantized. The log-band SNR estimates are scaled and adjusted by the total noise energy in each band to avoid SNR dependent fluctuations in the output signal energy [84].

The channel SNRs are employed to make a voice activity decision. Non-speech frames are used to update noisy background statistics. Regardless of the VAD decision, the background noise is updated after 35 speech frames. Sudden changes in environmental noise can thus be handled.

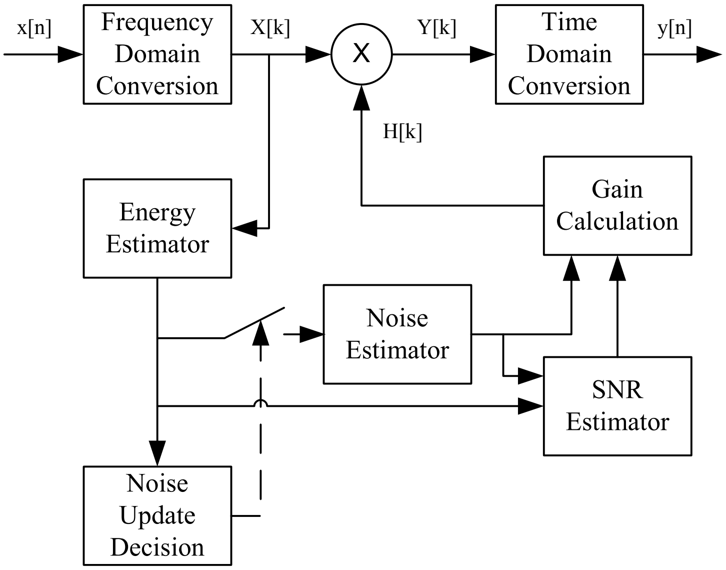


Fig. C.1 Overview of the EVRC algorithm [5]

Channel gains are calculated as a function of the channel SNR increasing linearly within the range of −13 to 0 dB.

References

[1] G. H. Golub and C. F. V. Loan, *Matrix Computations.* John Hopkins, 3rd ed., 1996.

[2] H. Anton, *Elementary Linear Algebra — Abridged Version.* Wiley, 7th ed., 1994.

[3] P. Sorqvist, P. Handel, and B. Ottersten, "Kalman filtering for low distortion speech enhancement in mobile communication," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 2, (Munich, Germany), pp. 1219–1222, Apr. 1997.

[4] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[5] T. V. Ramabadran, J. P. Ashley, and M. J. McLaughlin, "Background noise suppression for speech enhancement and coding," in *Proc. IEEE Workshop on Speech Coding For Telecommunications*, (Pocono Manor, Pennsylvania), pp. 43–44, Sept. 1997.

[6] J. S. Collura, "Speech enhancement and coding in harsh acoustic noise environments," in *Proc. IEEE Workshop on Speech Coding*, vol. 2, (Porvoo, Finland), pp. 162–164, May 1999.

[7] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. IEEE Workshop on Speech Coding*, (Porvoo, Finland), pp. 165–167, May 1989.

[8] M. Kuropatwinski, D. Leckschat, K. Kroschel, and A. Czyzewski, "Integration of speech enhancement and coding techniques," in *Proc. IEEE Workshop on Speech Coding*, (Porvoo, Finland), pp. 168–170, May 1999.

[9] D. Kim, Y. Park, I. Kim, and S. Park, "The effect of the speech enhancement algorithm for the sensorineural hearing impairment listener," in *Proc. Twentieth Annual Int. Conf. IEEE Eng. in Med. and Bio. Soc.*, vol. 6, (Piscataway, New Jersey), pp. 3150–3153, Oct. 1998.

[10] N. A. Whitmal, J. C. Rutledge, and L. A. Wilber, "An evaluation of wavelet-based noise reduction for digital hearing aids," in *Proc. Nineteenth Annual Int. Conf. IEEE Eng. in Med. and Bio. Soc.*, vol. 5, (Salt Lake City, Utah), pp. 4005–4008, Oct. 1997.

[11] D. O'Shaughnessy, P. Kabal, D. Bernardi, L. Barbeau, C.-C. Chu, and J.-L. Moncet, "Applying speech enhancement to audio surveillance," in *Proc. IEEE Int. Carnahan Conf. on Crime Countermeasures*, (Lexington, KY), pp. 69–71, Oct. 1988.

[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[13] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, (Washington, DC), pp. 208–211, Apr. 1979.

[14] H. L. V. Trees, *Detection, Estimation, and Modulation : Part I - Detection, Estimation and Linear Modulation Theory.* John Wiley and Sons, Inc., 1st ed., 1968.

[15] A.-J. D. Veen, E. F. Deprettere, and A. L. Swindlhurst, "Subspace-based signal analysis using singular value decomposition," *Proc. IEEE*, vol. 81, pp. 1277–1308, Sept. 1993.

[16] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 251–266, July 1995.

[17] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, pp. 45–57, Feb. 1991.

[18] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophys. J. R. Astr. Soc.*, vol. 33, pp. 347–366, 1973.

[19] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. AP-34, pp. 276–280, Mar. 1986.

[20] D. W. Tufts, R. Kumaresan, and I. Kirsteins, "Data adaptive signal estimation by singular value decomposition of a data matrix," *Proc. IEEE*, vol. 70, pp. 684–685, June 1982.

[21] B. De Moor, "The singular value decomposition and long and short spaces of noisy matrices," *IEEE Trans. Signal Processing*, vol. 41, pp. 2826–2838, Sept. 1993.

[22] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in colored noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, (Seattle, WA), pp. 377–380, May 1998.

[23] J. Huang and Y. Zhao, "A DCT-based fast signal subspace technique for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 747–751, Nov. 2000.

[24] A. Rezayee and S. Gazor, “An adaptive KLT approach for speech enhancement,” *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.

[25] B. Yang, “Projection approximation subspace tracking,” *IEEE Trans. Signal Processing*, vol. 43, pp. 95–107, Jan. 1995.

[26] U. Mittal and N. Phamdo, “Signal/noise KLT based approach for enhancing speech degraded by colored noise,” *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 159–167, Mar. 2000.

[27] F. Jabloun and B. Champagne, “On the use of masking properties of the human ear in the signal subspace speech enhancement approach,” in *Int. Workshop on Acoustic Echo and Noise Control*, (Darmstadt, Germany), Sept. 2001.

[28] G. A. Soulodre, *Camera Noise from Film Soundtracks.* Ph.D. thesis, McGill University, Department of Electrical Engineering, Nov. 1998.

[29] N. Virag, “Signal channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.

[30] D. O’Shaughnessy, *Speech Communications — Human and Machine.* IEEE Press, 2nd ed., 2000.

[31] J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals.* Macmillan Publishing Company, 1st ed., 1993.

[32] S. Haykin, *Adaptive Filter Theory.* Prentice Hall, 3rd ed., 1996.

[33] R. J. McAulay and T. F. Quatieri, “Speech analyis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[34] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis.* Elsevier, 1st ed., 1995.

[35] E.-B. Fgee, W. J. Phillips, and W. Robertson, “Comparing audio compression using wavelets with other audio compression schemes,” in *Proc. IEEE Int. Canadian Conf. on Electrical and Computer Engineering*, vol. 2, (Edmonton, Alberta), pp. 698–701, May 1999.

[36] N. M. Hosny, S. H. El-Ramly, and M. H. El-Said, “Novel techniques for speech compression using wavelet,” in *Proc. Eleventh Int. Conf. on Microelectronics*, (Kuwait City, Kuwait), pp. 225–229, Nov. 1999.

[37] I. T. Jolliffe, *Principal Component Analysis.* Springer Series in Statistics, Springer-Verlag, 1st ed., 1986.

[38] A. Dür, "On the optimality of the discrete Karhunen-Loève expansion," *SIAM J. Control Optim.*, vol. 36, pp. 1937–1939, Nov. 1998.

[39] Y. Hua and W. Liu, "Generalized Karhunen-Loève transform," *IEEE Signal Processing Letters*, vol. 5, pp. 141–142, June 1998.

[40] J. H. Wilkinson, *The Algebraic Eigenvalue Problem.* Clarendon Press, 1st ed., 1965.

[41] J. Rissanen, "Modeling by short data description," *Automatica*, vol. 14, pp. 465–471, Sept. 1978.

[42] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, Sept. 1986.

[43] A. Kavčić and M. Srinivasan, "The minimum description length principle for modeling recording channels," *IEEE J. on Selected Areas in Comm.*, vol. 19, pp. 719–729, Apr. 2001.

[44] T. W. Anderson, "Asymptotic theory for principal component analysis," *IEEE Trans. Speech and Audio Processing*, vol. 34, pp. 122–148, Mar. 1963.

[45] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 33, pp. 387–392, Apr. 1985.

[46] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, vol. 15 of *Series in Computer Science.* World Scientific, 1st ed., 1989.

[47] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, Mar. 1978.

[48] N. Merhav, "The estimation of model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1109–1114, Sept. 1989.

[49] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *Automatica*, vol. 35, pp. 1014–1019, Sept. 1989.

[50] B. C. J. Moore, *An Introduction to the Psychology of Hearing.* Academic Press, 4th ed., 1997.

[51] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models.* Springer, 2nd ed., 1998.

[52] B. C. J. Moore, J. I. Alcàntra, and T. Dau, "Masking patterns for sinusoidal and narrow-band noise maskers," *J. Acoust Soc. Am.*, vol. 104, pp. 1023–1038, Aug. 1998.

[53] R. P. Hellman, “Asymmetry of masking between noise and tone,” *Perception & Psychophysics*, vol. 11, pp. 241–246, Mar. 1972.

[54] W. Jesteadt, S. Bacon, and J. Lehman, “Forward masking as a function of frequency, masker level, and signal delay,” *J. Acoust Soc. Am.*, vol. 71, pp. 950–962, Apr. 1982.

[55] A. J. Oxenham and B. C. J. Moore, “Modeling the additivity of nonsimultaneous masking,” *Hearing Research*, vol. 80, pp. 105–118, None 1994.

[56] J. O. Pickles, *An Introduction to the Physiology of Hearing.* Academic Press, 2nd ed., 1988.

[57] T. J. Lynch III, W. T. Peake, and V. Nedzelnitsky, “Input impedance of the cochlea in cat,” *J. Acoust Soc. Am.*, vol. 72, pp. 108–130, July 1982.

[58] *Method for objective measurements of perceived audio quality*, Recommendation ITU-R BS.1387, International Telecommunication Union, July 1999.

[59] E. Terhardt, “Calculating virtual pitch,” *Hearing Research*, vol. 1, pp. 155–182, 1979.

[60] H. Fletcher, “Auditory patterns,” *Revs. Modern Phys.*, vol. 12, pp. 47–65, Jan. 1940.

[61] M. R. Schroeder, B. S. Atal, and J. L. Hall, “Optimizing digital speech coders by exploiting masking properties of the human ear,” *J. Acoust Soc. Am.*, vol. 66, pp. 1647–1652, Dec. 1979.

[62] E. Terhardt, G. Stoll, and M. Seewann, “Algorithm for extraction of pitch and pitch salience from complex tonal signals,” *J. Acoust Soc. Am.*, vol. 71, pp. 679–688, Mar. 1982.

[63] D. M. Green, “Additivity of masking,” *J. Acoust Soc. Am.*, vol. 41, pp. 1517–1525, Jan. 1967.

[64] R. A. Lufti, “Additivity of simultaneous masking,” *J. Acoust Soc. Am.*, vol. 73, pp. 262–267, Jan. 1983.

[65] J. D. Johnston, “Transform coding of audio signals using perceptual noise criteria,” *IEEE J. on Selected Areas in Comm.*, vol. 6, pp. 314–323, Feb. 1988.

[66] S. J. Godsil and P. J. Rayner, *Digital Audio Restoration.* Springer, 1st ed., 1998.

[67] J. K. Thomas, L. L. Scharf, and D. W. Tufts, “The probability of a subspace swap in the SVD,” *IEEE Trans. Signal Processing*, vol. 43, pp. 730–736, Mar. 1995.

[68] M. Hawkes, A. Nehorai, and P. Stoica, “Performance breakdown of subspace-based methods: Prediction and cure,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 6, (Salt Lake City, Utah), pp. 4005–4008, May 2001.

[69] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, “Speech enhancement based on audible noise suppression,” *IEEE Trans. Speech and Audio Processing*, vol. 5, pp. 497–514, Nov. 1997.

[70] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing — Principles, Algorithms and Applications*. Pretice Hall, 3rd ed., 1996.

[71] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[72] *IEEE Recommended Practice for Speech Quality Measurements*, Standards Publication No. 297, Institute of Electrical and Electronics Engineers, Sept. 1969.

[73] “Signal Processing Information Base.” Content at http://spib.rice.edu/spib/spib.html, URL current as of Dec. 2001.

[74] *Subjective Performance Assessment of Telephone-Band Wideband Digital Codecs*, Recommendation ITU-T P.830, International Telecommunication Union, Feb. 1996.

[75] *Objective Measurement of Active Speech Level*, Recommendation ITU-T P.56, International Telecommunication Union, Mar. 1993.

[76] R. W. Berry, “Speech-volume measurements on telephone circuits,” *Proc. IEEE*, vol. 118, pp. 335–338, Feb. 1971.

[77] P. Kabal, “Measuring speech activity,” tech. rep., McGill University, Aug. 1999.

[78] *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, Recommendation ITU-R BS.1116-1, International Telecommunication Union, Oct. 1997.

[79] *Perceptual evaluation of speech quality (PESQ)*, Recommendation ITU-T P.862, International Telecommunication Union, Feb. 2001.

[80] V. Sánchez, P. García, A. M. Peinado, J. C. Segura, and A. J. Rubio, “Speech-volume measurements on telephone circuits,” *IEEE Trans. Signal Processing*, vol. 43, pp. 2631–2641, Nov. 1995.

[81] *Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s — Part 3: Audio*, IS11172-3 1993, ISO/IEC, JTC1/SC29/WG11, Apr. 1993.

[82] M. Avriel, *Nonlinear Programming: Analysis and Methods.* Series in Automatic Computation, Prentice-Hall, 1st ed., 1976.

[83] *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Systems*, Interim Standard IS-127, International Telecommunication Union, Jan. 1996.

[84] M. C. Recchione, “The enhanced variable rate coder: Toll quality speech for cdma,” *Int. J. Speech Technol*, vol. 2, pp. 305–315, May 1999.