INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Adaptive Methods for Removing Camera Noise from Film Soundtracks

by

Gilbert A. Soulodre

A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Department of Electrical Engineering
McGill University
Montreal, Canada
November 1998

ABSTRACT

One of the fundamental problems in signal processing is to enhance a signal which has been corrupted by an additive noise. In this thesis, the problem of alleviating the effects of camera noise corrupting the dialog of a film soundtrack is examined. Two methods of noise reduction are investigated: adaptive noise cancellation with a synthesized reference signal, and spectral subtraction. It is found that, due to the relatively low correlation between successive camera noise pulses, the adaptive noise cancellation approach is not effective at reducing camera noise. The spectral subtraction method is shown to reduce camera noise, but the process creates audible artifacts which can be very disturbing to the listener. To overcome this, new methods are proposed for reducing musical noise and time aliasing effects. The use of subbands and sub-frames is shown to significantly improve the performance of the spectral subtraction algorithm by providing a better match of the noise reduction process to the noise. The performance is further improved by incorporating a perceptual model into the spectral subtraction algorithm. The use of subbands, sub-frames, and a perceptual model allows the amount of processing applied to the signal to be minimized which in turn reduces the level of any artifacts which may result from the noise reduction process. The results of a formal subjective test demonstrate the improved performance of the new noise reduction algorithm.

RÉSUMÉ

De tous les problèmes rencontrés en traitement de signal, un des plus fondamentaux est l'amélioration d'un signal détérioré par un bruit additif. Cette thèse considère le problème de réduction de bruits provenant d'une ciné-caméra et se retrouvant dans la bande sonore d'un film et donc affectant le dialogue. Deux approches sont étùdiées pour réaliser cette réduction de bruit, soit l'annulation adaptative du bruit par la méthode des moindres carrées et utilisant un signal de référence synthétisé, ainsi que la soustraction spectrale. Les résultats démontrent que l'approche d'annulation adaptative du bruit utilisant un signal de référence synthétisé n'est pas efficace pour réduire le bruit de caméra, ceci étant du à la corrélation relativement faible entre les impulsions consécutives du bruit. L'approche de soustraction spectrale offre une réduction considérable du bruit de caméra, mais des artifices perceptibles et très perturbant pour l'auditeur en résultent. Pour éviter ceci, de nouvelles approches de réduction de bruit musical et d'effets de repliement (aliasing) temporel sont proposées. L'utilisation de sous-bandes et de sous-trames résulte en une amélioration importante de la performance de l'approche de soustraction spectrale en créant une meilleur correlation entre le bruit et the processus. Le rendement est d'autant plus amélioré en insérant un model perceptuel dans l'algorithme de soustraction spectrale. L'utilization de sous-bandes, de sous-trames, et du model perceptuel permet de minimiser le degré de traitement du signal qui en plus minimise les artifacts provenant de la réduction de bruit. Les résultats d'un test subjectif formel démontrent l'amélioration de la performance de l'algorithme de réduction de bruit.

ACKNOWLEDGMENTS

Although only one name appears on the title page of this dissertation, many people should be acknowledged for their contributions which either directly or indirectly made this work possible.

First, I would like to thank my thesis advisor, Professor Peter Kabal for his friendship and guidance in seeing this work through to completion.

A thesis on the topic of camera noise would not be much of a thesis without recordings of camera noise. I would like to thank David Simpson of the National Film Board of Canada for providing me with a camera, film, and a skilled camera operator. I would also like to thank J.P. Vialard for his help in coordinating the measurements and for providing me with access to the NFB's mix-down studio.

All of the computer simulations performed in this thesis were done with the rather meager computing power in my home. I would like to thank a good friend, John Oldfield, who's virtuosity as a hacker provided me with various bits and pieces to squeeze the maximum performance out of my computer.

I am indebted to my colleagues at the Communications Research Centre for their encouragement, support, and valuable discussions. So long Ted.

I would like to thank Marie for sleeping through the night so that her dad could work through the night, and David, who's birth provided me with some much needed motivation.

Finally, I would like to thank my wife, Eleanor who; urged me to start my Ph.D., supported me throughout my courses and thesis work, and coaxed me towards its completion. Her personal sacrifice during this time is immeasurable.

This thesis is dedicated to the memory of my brother
who first got me interested in audio and electronics.

Robert Pierre Joseph Soulodre

1950-1998

In the time of my confession, in the hour of my deepest need.
When the pool of tears beneath my feet flood every newborn seed.
There's a dyin' voice within me reaching out somewhere,
Toiling in the danger and the morals of despair.
Don't have the inclination to look back on any mistake.
Like Cain, I now behold the chain of events that I must break.
In the fury of the moment I can see the Master's hand
In every leaf that trembles, in every grain of sand.

Oh, the flowers of indulgence and the weeds of yesteryear,
Like criminals they have choked the breath of conscience and good cheer.
The sun beat down upon the steps of time to light the way
To ease the pain of idleness and the memory of decay.
I gaze into the doorway of temptation's angry flame
And every time I pass that way I always hear my name.
Then onward in my journey I come to understand
That every hair is numbered like every grain of sand.

I have gone from rags to riches in the sorrow of the night
In the violence of a summer's dream, in the chill of a wintry night.
In the bitter dance of loneliness fading into space,
In the broken mirror of innocence on each forgotten face.
I hear the ancient footsteps like the motion of the sea
Sometimes I turn, there's someone there, other times it's only me.
I am hanging in the balance of the reality of man
Like every sparrow fallen, like every grain of sand.

R. Zimmerman

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

This thesis addresses one aspect of the general problem of enhancing a signal which has been corrupted by an additive noise. This problem arises in applications ranging from removing noise from speech signals in a telephone system, to detecting sonar signals amidst the ambient noise of the ocean, to enhancing fetal electrocardiograms. In this thesis, we examine the problem of enhancing a speech signal which has been corrupted by a repetitive or cyclical noise source. That is, it is assumed that certain characteristics of the interfering noise repeat over time.

While the research described in this thesis focuses on the specific application of removing camera noise from film soundtracks, the results are readily extendible to other applications which require that a signal be enhanced in the presence of a repetitive noise. For example, many mechanical devices (e.g. motors, generators, cooling fans, printing presses, propeller blades, etc.) produce repetitive acoustic noises which can corrupt a desired acoustic signal (e.g. speech, music, sonar). Also, there are many sources of cyclical electrical noise (e.g. car ignition noise, interference from electric motors, switching power supplies, etc.) which can corrupt other electrical signals. Therefore, any signal which has been corrupted by a repetitive noise source can potentially be enhanced using the methods described in this thesis.

As a result of one of the noise reduction methods investigated in this thesis, a mathematical and subjective comparison of two auditory masking models was conducted, and new enhancements to the models were proposed. This work has implications for many applications beyond the field of noise reduction.

1.1 Description of the Problem of Camera Noise in Film Soundtracks

This thesis examines the problem of camera noise in film soundtracks and investigates potential schemes for reducing this noise. The soundtrack of a motion picture consists of a mix of audio recordings of music, sound effects, and speech. The music segments of a soundtrack are invariably recorded in the highly controlled acoustic environment of a recording studio. Similarly, sound effects are often taken from a pre-recorded library of sounds, or if they are not available from such a library, they are created and recorded on a foley stage. The foley stage is another acoustically controlled environment

designed specifically for the task of creating sound effects for films. Since neither the music nor the sound effects are recorded at the time of filming, they are not affected by camera noise. The dialogue however, is recorded at the time of filming and it is here where the problem of camera noise arises.

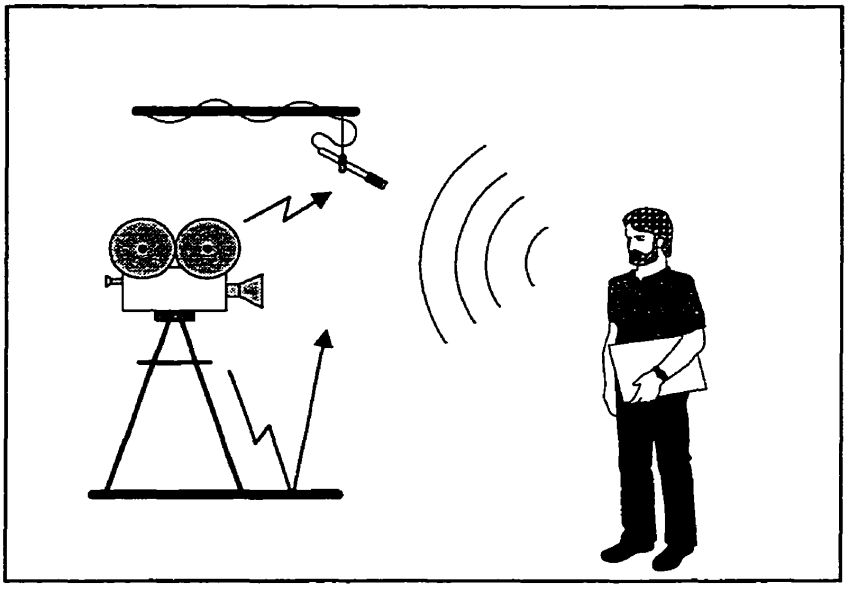


Figure 1.1 Illustration of a typical filming scenario.

The problem of camera noise corrupting the dialogue recordings can be described with the help of Figure 1.1. The figure depicts a typical filming scenario with an actor standing before a camera reciting his dialogue. Above the camera is a microphone which is used to record the actor's voice and any other assorted sounds that the actor might make (e.g. coughing, rustling paper, typing, footsteps etc.). The microphone is typically placed as close to the actor as is possible without appearing within the view of the camera.

Simply stated, the fundamental problem is that due to its mechanical workings, the camera produces an acoustical noise which is picked up by the microphone (as depicted by the arrows in Figure 1.1) along with the speech signal. The noise of the camera is superimposed onto the soundtrack along with the voice of the actor. The loudness of this noise relative to the level of the actor's voice can vary significantly due to: the type of camera; the type of microphone employed; the acoustic characteristics of the room; and the relative positions of the camera, the actor and the microphone. It is possible to limit the loudness of the camera noise to some extent by carefully controlling some or all of these variables. However, this reduction in the level of the noise may occur at the expense of limiting some other technical or artistic aspect of the filming process. For ex-

ample, unidirectional microphones which are more sensitive to sounds arriving from the front of the microphone than from its rear, can be employed to record the dialogue. By placing the camera to the rear of a unidirectional microphone, it may be possible to reduce the amount of camera noise picked up by the microphone. However, it should be noted that the camera noise reaches the microphone via many paths. That is, the camera radiates noise in all directions and the noise will reflect off of the various surfaces (e.g. floor, walls, furniture, etc.) within the room and will reach the microphone at different times, from various directions and with different amplitudes. Therefore, the reverberation of the room may limit the effectiveness of this approach. If the level of the camera noise is relatively low in comparison to the dialogue then, as an alternative to reducing the noise, it may be possible to mask it with background music or sound effects.

Despite these means of limiting or masking the camera noise in the dialogue recordings, it is still very common for the camera noise to be audible to some degree. This is particularly true for IMAX™ films* since the larger cameras used to make these films are inherently noisier. Even small amounts of camera noise may distract the viewer and destroy the film's illusion of reality, and therefore, any audible camera noise is generally considered to be unacceptable [1].

For those cases in which the level of the camera noise is sufficiently high so as to be detectable, the actor's dialogue must be re-recorded (dubbed) in an acoustically controlled environment after the filming has occurred. This process is known as automatic dialog replacement (ADR). In ADR, the actors recite their dialogue while watching an image of their (previously filmed) performance and listening to the noisy version of their dialogue. While ADR completely eliminates the problem of camera noise, it is an undesirable solution since it adds significant costs to the production of a film and, because the actor must now worry about remaining synchronized with the image, it typically compromises his performance [1,2]. ADR is used regularly during the making of most movies in order to overcome the problem of camera noise. Therefore, a method for removing camera noise without adversely affecting the underlying speech signal would be of significant benefit to the film making process [3,4]. A review of the pertinent scientific literature and discussions with individuals in the film industry indicate that no such method presently ex-

* The IMAX corporation produces a specialized type of motion picture using a very large screen format which encompasses the viewer's peripheral field of view. To retain resolution and picture quality, IMAX cameras require a sophisticated transport mechanism.

ists. However, the IMAX Corporation, whose cameras pose a greater noise problem due to their large size, did conduct some research on this topic [5].

It is reasonable to question the value of a system which could successfully reduce camera noise. In the making of a film, the cost of ADR is typically on the order of US$50,000. Given the hundreds of films made each year, one can see that literally millions of dollars are spent annually on ADR. This is despite the methods which currently exist for limiting camera noise (see Chapter 2).

1.2 Description of Camera Noise

It is useful to have in mind an idea of the sound of camera noise. To this end, it is instructive to describe the mechanisms which combine to produce the acoustic noise in a motion picture camera.

A motion picture camera is an intricate mechanical device composed of many moving parts. The film is transported from the supply reel, through the camera to the take-up reel by means of the sprocket holes which line the film. The sprocket system is used to ensure that the film is correctly aligned with the camera's shutters and lens. With a frame of the film correctly aligned, the shutters open and close to briefly expose the film, and the film is then moved to the next frame. This process is repeated 24 times every second. The rate (film rate) of 24 frames per second was chosen to provide sufficient visibility of lip movement when sound was introduced to motion pictures [6]. The camera noise is heard as a series of clicks or pulses (actually, noise bursts) occurring at a rate of 24 times per second.

The reader is probably familiar with the sound of a motion picture projector. Given that the camera and projector have many similar components, the noise of the camera is quite similar to that produced by a projector. Examples of camera noise can be heard on the compact disc accompanying the thesis (see Chapter 8).

1.3 Requirements of a Noise Reduction Scheme

In order for any camera noise reduction scheme to be fully acceptable, there are several requirements which it must meet. Of course, a successful noise suppression technique must reduce the camera noise such that any residual noise will not be perceptible to the audience. Also, the process must not adversely affect the quality of the underlying speech

signal. This implies that the restored speech signal must be of very high quality and that no audible artifacts can be introduced as a result of the noise reduction process.

Given the manner in which dialogue is traditionally recorded, it is highly desirable that the noise suppression technique be a single-input system. That is, the technique should be able to process the corrupted (noisy) signal without the benefit of an additional recording of the isolated camera noise. While this is a very severe restriction in that it eliminates the use of certain approaches for noise suppression, it is unlikely that any scheme would gain widespread acceptance unless it meets this fundamental requirement [3,4,5]. Moreover, a successful single-input noise suppression scheme could also be used for the restoration of older films, thus greatly extending its usefulness.

Since a successful noise reduction scheme would be used as part of an artistic process (i.e. the making of a soundtrack for a film), the key parameters which control the performance of the process must be identified and put under the user's control. Finally, while it is not absolutely necessary for the noise reduction to occur in real-time, from a practical point of view, the process must operate with reasonable speed.

1.4 Noise Reduction Techniques

In describing the problem of camera noise it was seen that the task of removing camera noise from a film soundtrack consists primarily of extracting a speech signal from noise. Separating a desired speech signal from an undesired signal is an important and common problem in signal processing. There has been a significant amount of research devoted to this topic, although primarily in the context of telephony, speech compression, speech recognition and military voice communications. In this section, methods of noise reduction based on adaptive filtering methods and spectral subtraction are considered briefly. Some of the benefits and limitations of each approach are also addressed.

1.4.1 Adaptive Noise Cancellation Algorithms

A commonly used method for reducing noise in speech signals is the adaptive noise cancellation (ANC) technique. In its basic form, ANC uses two inputs: a primary input and a reference input. In the present application, the primary input corresponds to the noisy speech (recorded at the primary microphone), while the reference input would consist of a recording of the camera noise alone. In practice, the reference input would be obtained by placing a second receiver (reference microphone) next to the camera and recording the camera noise at the same time that the dialogue is being recorded. To understand how the

adaptive noise cancellation method works, consider the propagation of the noise from the camera to the (primary) microphone as depicted in Figure 1.1. The sound from the camera first reaches the microphone by the direct path between them. Due to the reverberation of the room however, this is followed by a plethora of reflections arriving at different times and with different amplitudes. The ANC method works by estimating this complex acoustic response from the camera to the primary microphone.

Given this estimate, the reference input (i.e. the camera noise) is processed to produce an approximation of the camera noise as it would appear at the primary microphone. This approximation of the camera noise is then subtracted from the primary input signal, thus reducing the noise and leaving a noise-reduced recording of the speech signal. The estimation of the acoustic path is often done by using the least-mean square (LMS) adaptive algorithm which adapts on an iterative basis. A comprehensive introduction to ANC and its many applications, as well as a derivation of the LMS algorithm can be found in the classic paper by Widrow *et al.* [7].

While ANC can be quite effective in many applications, there are certain basic limitations which must be considered. Since the LMS algorithm iteratively derives its estimate of the acoustic path, the rate at which the algorithm adapts to produce this estimate must be considered. Clearly, it is desirable for the algorithm to adapt as quickly as possible. However, the speed with which the LMS algorithm can adapt is limited by the opposing requirements for the algorithm to remain stable and the need for an accurate estimate. Performance of an ANC system can also be compromised if some of the desired speech signal leaks into the reference input signal (i.e. is recorded by the reference microphone). Variations to the basic LMS algorithm and an analysis of the factors which limit its performance (adaptation rate, accuracy of the estimate, stability, etc.) can be found in [8,9,10,11]. An important variant of the LMS algorithm is the normalized LMS algorithm, which can provide superior performance when dealing with impulse-type noise (such as camera noise).

An area of research which is closely related to ANC is the problem of blind signal separation [75,81,88,89,91]. Blind signal separation can be viewed as a generalization of the ANC method which attempts to overcome many of the limitations inherent to ANC.

It was originally believed when this research work began that LMS-based ANC was the obvious choice of methods for reducing camera noise. There are several factors however which make the use of this method less appealing for this application. The most important factor is that ANC is a two-input, rather than a single-input scheme. A single-

input scheme, of course, was one of the primary criteria for an acceptable noise reduction scheme. In some instances, when the noise is repetitive and somewhat predictable, it may be possible to synthesize the reference signal rather than record it directly. This effectively creates a single-input ANC system. For this approach to reduce camera noise requires that the individual pulses of the camera noise be sufficiently similar to each other so that a representative reference signal can be derived. Unfortunately as will be seen in Chapter 4, the individual pulses are not similar enough to yield a sufficient degree of noise reduction using this approach.

1.4.2 Methods Based on Spectral Subtraction Techniques

At about the same time that adaptive noise cancellation was first being developed, the technique of spectral subtraction was proposed by Weiss *et al.* [12] and by Boll [13,14]. The process was originally intended for military applications in an attempt to improve the intelligibility of speech under extreme noise conditions. For example, spectral subtraction was used to try to improve voice communications in the cockpits of jet fighter aircraft and helicopters [13,15]. Interestingly, tests showed that the spectral subtraction technique did not provide any improvement in intelligibility [13,16,17]. It did however provide a perceived improvement to the quality of the processed speech signal, and it is in this context that spectral subtraction is examined as a potential means of reducing camera noise.

In a manner similar to ANC, spectral subtraction forms an approximation of the noise signal and then subtracts this estimate from the noisy speech signal. However, spectral subtraction uses a much less precise approximation of the noise signal than ANC. More precisely, an audio signal can be described in terms of its combined spectral *magnitude* and *phase*. The adaptive noise cancellation scheme described earlier requires an accurate determination of both of these parameters, whereas the spectral subtraction process only estimates the spectral magnitude and effectively ignores the phase. Ignoring the phase portion of the noise causes corresponding errors in the phase portion of the processed speech. However, these errors are usually unimportant since the ear has been found to be relatively insensitive to the phase portion of speech signals [18,19].

This simplified approximation of the noise signal provides several advantages. For example, unlike the ANC scheme, the spectral subtraction algorithm does not suffer from conflicts between stability and adaptation rate. More importantly however, the simplified approximation allows spectral subtraction to be a single-input process thus making it

suitable for the task of reducing camera noise. Spectral subtraction relies on the assumption that the spectral magnitude of the noise during gaps in the speech is the same as during speech intervals. Therefore, the spectral subtraction scheme derives its noise estimate directly from the recording of the noisy speech during the intervals where there is no speech activity.

Due to its crude characterization of the noise source, the spectral subtraction process can produce many audible artifacts which are sometimes more disturbing to the listener than the original noise. The artifacts become more audible as more aggressive processing is applied to the noisy speech signal. As the level of the camera noise increases, more aggressive processing must be applied to sufficiently reduce this noise, and thus the resulting artifacts become more audible.

Spectral subtraction is actually a name given to a family of algorithms which are variations of a fundamental technique. The various algorithms differ primarily in how they form their estimate of the noise signal. They effectively provide a trade-off between the amount of noise suppression achieved and the level of the resulting artifacts. Perhaps the most disturbing artifact, and certainly the one which has received the most attention, is *musical noise*. Musical noise can be described as a sequence of short-lived tones occurring at random times and frequencies. A significant portion of the research into spectral subtraction has been devoted to studying ways of limiting and suppressing musical noise. Vaseghi and Frayling-Cork [20] proposed a 'survival algorithm' for removing musical noise which is based on examining the amplitude and duration of the tones that make up the musical noise. Cappé [21] has shown how the variant of the spectral subtraction algorithm proposed by Ephraim and Mahler [22] provides noise reduction without creating musical noise. However, this version of the spectral subtraction algorithm does not completely eliminate the interfering noise signal and is therefore not applicable to the camera noise problem.

An interesting variation to the spectral subtraction algorithm was investigated by Tsoukalas *et al.* [120,130,23] wherein a model of the human auditory system is incorporated into the system. The auditory model is used to determine which portion of the noise is audible and which is being masked by the desired signal. The spectral subtraction algorithm then removes only that portion of the noise which is audible. This approach is reported to significantly reduce artifacts such as musical noise [120,121,130,131,132].

Spectral subtraction has been applied mainly to reducing high level noise in voice communication systems and as a pre-processor to speech compression and speech recog-

nition systems [24,25,26,14,27,28,29]. In these applications, where noise conditions can be quite severe, artifacts resulting from the processing may be acceptable provided that communications (or recognition) are improved. Spectral subtraction has also been used to reduce the background noise (hiss) in old gramophone recordings prior to being transferred to compact disc [20,30]. In this application spectral subtraction was found to work successfully if the level of the background noise is sufficiently far below (>30dB) the level of the music signal. This is a much less severe noise condition than found in some voice communication applications. Also, in the gramophone restoration application it is acceptable to merely reduce the level of the background noise without making it inaudible. This is in contrast to the camera noise problem where the noise must be rendered inaudible. As with the removal of camera noise, audible artifacts are not acceptable when restoring gramophone recordings.

The problem of removing camera noise shares features of both the voice communication and gramophone restoration applications. While the level of the camera noise is expected to be nearer to the noise levels found in communications systems, the processed speech signal must be of the same high quality demanded when restoring gramophone recordings. Therefore, removing camera noise makes for a unique and challenging noise reduction problem in that two opposing demands must be addressed. In this thesis it will be shown that spectral subtraction can be successfully used to remove camera noise in film soundtracks. This is achieved primarily by taking advantage of specific characteristics of the camera noise which allow the amount of processing applied to the noisy signal to be time and frequency dependent. By matching the noise reduction algorithm to the noise, the amount of processing applied to the signal can be reduced which in turn reduces the level of any residual artifacts. The use of a perceptual model in the spectral subtraction algorithm builds on this philosophy. By removing only those portions of the noise which are audible, the amount processing applied to the signal is reduced and thus the levels of the artifacts are also reduced.

1.5 Original Contributions

Based on a review of the pertinent literature and discussions with individuals in the film industry, this thesis appears to constitute the first comprehensive investigation into the use of adaptive signal processing methods for reducing camera noise in film soundtracks. The results of the thesis provide a successful single-input approach for removing camera noise while minimizing any audible effects on the underlying speech signal. As such,

these results point the way towards a hardware and/or software implementation which could be used both in the making of new films and for the restoration of older films.

An important aspect of the thesis work was the careful measurement and characterization of the acoustical and statistical properties of the camera noise. This was done for several cameras and the factors which cause variations in the camera noise were identified. It was shown that the camera behaves as a distributed noise source. This has important implications for the possible success of ANC-based noise reduction schemes. A mathematical model of camera noise was developed which differentiates between the periodic and cyclical random components of camera noise. The model was extended to include the inter-pulse jitter in the timing of the periodic component. The information derived in this chapter was not available in the scientific literature and will serve as a useful foundation for other researchers who may wish to address the problem. Moreover, the recordings form a valuable database which can be used by researchers to develop and evaluate other potential schemes for reducing camera noise.

Significant effort was given to investigating the use of ANC-based (and blind signal separation) techniques for reducing camera noise, and it was shown that this approach is not likely to yield a high degree of noise reduction due to the distributed nature of the camera noise. It was shown that the maximum amount of noise reduction is limited to about 15 dB which is insufficient for this application. These findings provide an explanation of why other (IMAX) attempts at using ANC for reducing camera noise failed.

A variation to ANC using a synthesized reference signal was proposed. Jitter in the timing of the camera noise was shown to limit the performance of this approach. A method for synchronizing the ANC to the camera noise was proposed and the resulting improvement in performance was demonstrated.

While the ANC approach was not successful at adequately suppressing camera noise, these negative results provide valuable information to future studies regarding the limitations of ANC in similar applications.

Several extensions and modifications to the traditional spectral subtraction algorithm were proposed which help to reduce some of the artifacts which can result from the process. These extensions are not restricted to the camera noise application and are useful to the general noise reduction problem.

- The zero-phase filter interpretation of the noise suppression equation was generalized and the effects of each of its parameters were analyzed.

- The various artifacts resulting from the noise reduction process were characterized and the cause of each artifact was related to one of two sources of error.
- The minimum spectral floor proposed by Berouti *et al.* [17] was extended to make the resulting noise floor more perceptually benign.
- The "survival algorithm" proposed by Vaseghi and Frayling-Cork [20] was extended to provide improved suppression of the musical noise.
- A new noise-overestimation parameter, based on the variance of the noise measured during the derivation of the noise estimate, was proposed.
- The use of zero-padded FFT's with truncation was proposed to reduce pre- and post- echoes (temporal smearing artifacts).
- An analysis/synthesis windowing operation was added to remove the discontinuities at the boundaries of the overlapping processing frames.

A general mathematical framework for a subband/sub-frame based spectral subtraction algorithm was derived which includes quadrature mirror analysis and synthesis filter banks. The possibility of aliasing when combining the subband signals (due to the spectral subtraction process) was highlighted and the implications for the filter bank design were considered. The use of subbands and sub-frames was shown to allow the noise reduction process to be matched to the characteristics of the noise, thus reducing the overall amount of processing. This follows from the general philosophy adopted in the thesis of minimizing the amount of processing applied to the signal in order to minimize the resulting artifacts. The approach was further generalized by using non-uniform sub-framing and issues regarding the appropriate choice of windows were addressed.

The need for frame synchronization when using spectral subtraction techniques in the presence of a cyclical interferer such as camera noise was identified. A simple means of obtaining frame synchronization was also proposed.

A significant amount of new work related to the topic of perceptual models was conducted in the research. The work provides direct benefit for camera noise reduction as well as general noise reduction applications. Moreover, the results are directly applicable to numerous other applications such as perceptual audio codecs which use a model of the human auditory system.

- A detailed mathematical comparison was made between the Zwicker and the Patterson-Moore models for simultaneous masking. This involved viewing the Zwicker model in a non-traditional manner (equivalent auditory filters in the linear frequency domain) and deriving generalized expressions for the Patterson-Moore auditory filters and excitation patterns.
- Significant differences were shown to exist between the two models, and these differences were shown to be both level and frequency dependent.
- The results of a psycho-physical study (in conjunction with Shlien) demonstrated that the ability of many listeners to discriminate variations in frequency is far superior to that predicted by the Zwicker model. This provides strong evidence of the need for a finer resolution basilar domain scale than is provided by the critical band model.
- A new analytic expression describing the filtering effects of the outer and middle ear was developed which recognizes the low frequency roll-off of the middle ear. A new complementary analytic expression for the internal noise floor of the auditory system was also derived.
- A new analytic expression was derived which predicts the amount of forward masking as a function of both frequency and level.
- A compression model for the addition of masking (both simultaneous and non-simultaneous) based on Humes and Jesteadt's modified power-law was integrated into the perceptual model to account for excess masking.
- The interaction (additional spreading in frequency) between the transform window and the auditory filters was demonstrated. The KBD window was shown to overcome many of the limitations inherent in "traditional windows". It was shown that the auditory filter model requires modifications in order to account for the frequency domain effects of the window.
- It was shown that the effects of the synthesis window are not included in the signal used by the perceptual model and thus, there is an inherent error (bias) in the predicted masking threshold. A method for resolving this matter was proposed and a window function which is appropriate for use with the KBD was derived.

A new perceptual model was developed based on: the auditory filters of Patterson and Moore modified to account for the effects of the transform window; the newly proposed outer and middle ear transfer functions; the newly proposed internal noise floor; the newly derived expression for forward masking; and the modified power-law for the addition of masking. The new perceptual model was incorporated into a subband/sub-frame based spectral subtraction algorithm. The algorithm included the new window preceding the perceptual model to account for the synthesis window, as well as the new method for estimating the clean signal.

- Subjective tests demonstrate the differences in the masking thresholds predicted by the Zwicker perceptual model versus the Patterson-Moore model. The results indicate that the Patterson-Moore model developed in this thesis provides better performance for the noise reduction application. The differences in the two perceptual models demonstrated in the thesis strongly suggest the need to reevaluate the use of Zwicker based models in perceptual audio codecs.
- A formal subjective test was conducted using the most rigorous and sensitive methods available. The results clearly demonstrate a significant improvement in the performance of the spectral subtraction algorithm due to the use of subbands and sub-frames, as well as the use of a perceptual model. The results also demonstrate that the methods developed in this thesis meet the requirements for a successful camera noise reduction system.

1.6 Outline of the Thesis

The thesis is divided into nine chapters. Chapter 2 provides a more detailed look at the problem of camera noise in film soundtracks and describes existing methods, and their limitations, for reducing its audibility. Also, previous (unsuccessful) research efforts to reduce camera noise are described. In Chapter 3, the properties of camera noise are characterized. Measurements of camera noise in both the time and frequency domains are described. These measurements provide valuable insights for investigating possible noise reduction schemes. Also, the distributed nature of camera noise is demonstrated and its implications on noise reduction are considered. A model of camera noise is proposed which allows it to be divided into two main (periodic and cyclical noise) components.

The fourth chapter examines the potential for using adaptive noise cancellation techniques to reduce camera noise. The chapter provides a review of the theory behind the LMS algorithm and some of its variants. The dependence of ANC on the coherence between the two input signals is demonstrated and its impact on the reduction of camera noise is discussed. The theory behind blind signal separation is reviewed and it performance is compared to ANC. The maximum amount of noise reduction attainable for camera noise using these methods is predicted. An ANC algorithm using a synthesized reference is proposed as a means of reducing one component of the camera noise. It is shown that, due to jitter in the camera noise, steps must be taken to keep the ANC process synchronized to the noise. The chapter ends with discussions and conclusions regarding the suitability of ANC methods for reducing camera noise.

Chapter 5 begins with an overview of spectral subtraction using Boll's method. Various spectral magnitude estimation methods are then analyzed and compared, and the performance of these methods is evaluated in the context of reducing camera noise. The artifacts resulting from spectral subtraction are described and several modifications and extensions are proposed which can help to reduce the severity of these artifacts. In Chapter 6, a spectral subtraction algorithm based on subbands and sub-frames is proposed to decompose the processing in the time-frequency plane. A mathematical framework is derived and it is shown that matching the noise reduction process to the camera noise can significantly improve the performance of the spectral subtraction algorithm. The use of a model of the auditory system to improve the performance of the spectral subtraction algorithm is considered in Chapter 7. Two well-known perceptual models are compared from a mathematical viewpoint, and several modifications are proposed. The new perceptual model is incorporated into the subband/sub-frame based spectral subtraction algorithm. The chapter concludes by examining how the transform window interacts with the perceptual model.

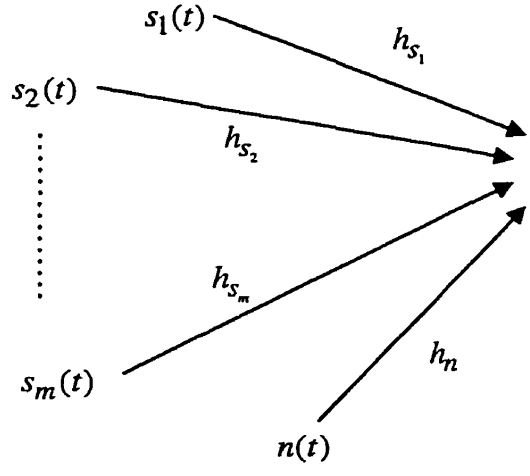
In Chapter 8 the various noise reduction schemes described and developed in the thesis are evaluated subjectively. The most promising schemes are evaluated in a formal subjective test. The results of the test provide a clear comparison of the performance of the various noise reduction schemes and demonstrate the improved performance due to the enhancements proposed in the thesis. Finally, Chapter 9 consists of a summary and discussion regarding the task of reducing camera noise in film soundtracks.

2. EXISTING NOISE REDUCTION METHODS

2.1 Introduction

The problem of camera noise was described in the previous chapter, where it was emphasized that any audible camera noise is generally considered unacceptable since it may destroy the sense of reality desired by a filmmaker. However, when watching a film, one is rarely aware of the sound of the camera and therefore, methods for preventing camera noise from corrupting the soundtrack must already exist. Indeed, there are several methods which currently exist for reducing the audibility of camera noise in film soundtracks and in this chapter, they are explored and their potential limitations are highlighted.

The methods outlined in Section 2.2-*Microphone Techniques* and Section 2.3-*Acoustic Barriers and Blimps* are employed at the time of filming and may by their nature impose limitations on creative aspects of the filming process. Section 2.4 describes an electronic (analog) signal processing device which, though not intended for this purpose, can be used to reduce camera noise to some extent. The main benefit of this approach is that, since it is a post-processing approach, no limitations are imposed at the time of filming. Dubbing or automatic dialogue replacement is described in Section 2.5 as the method of last resort. Although ADR entirely eliminates the problem of camera noise, it can be a rather costly solution and can have an impact on the final artistic quality of the soundtrack. Section 2.6 discusses informal results of several attempts to use modern signal processing methods to reduce camera noise.



$$y(t)=\sum_{i=1}^{m}\left[s_i(t)*h_{s_i}\right]+n(t)*h_n$$

Figure 2.1 Simplified model of the camera noise problem.

The problem of camera noise is reexamined in Figure 2.1. The figure shows the signal $y(t)$ received at a microphone which is a summation of m signals ($s_i(t)$, i=1,2,...m) and the camera noise $n(t)$. The m signals represent the various actors as well as any other sounds which are desired during the time of recording. Each of the m signals is convolved with a corresponding acoustic impulse response, h_{s_i}. Similarly, the camera noise is also convolved with an acoustic impulse response, h_n. Each of the impulse responses represents the acoustic path from a source to the microphone. The possible methods for reducing camera noise will be examined in the context of this simple model.

2.2 Microphone Techniques

Before discussing how microphone techniques might be used to reduce camera noise, consider some of the goals of the sound engineer when recording the dialogue for a film soundtrack. First, it must be recognized that the goals go far beyond merely making an intelligible recording of the dialogue. The recording must also capture the timbre, reverberation, and spatial characteristics associated with the sound of the actors' voices in the given acoustic environment [1]. Furthermore, the recording must also capture other acoustic events not associated with the dialogue. That is, other sounds such as background noises (e.g., passing cars, environmental sounds, etc.) and any incidental sounds made by the actors also need to be recorded at the time of filming. This implies that the use of highly directional microphones to focus-in on the actors' voices may be an inadequate approach and that one or more less directional microphones may be required.

When recording dialogue in a room, the microphone not only picks up the sound directly from the actor, but also picks up the many reflections from the surfaces within the room. These reflections have a variety of amplitudes and delays with respect to the direct sound, and it is the relation between the direct sound and these reflections which determines the acoustic character of the room. For example, in a small room (e.g. office, living room) where the reflecting surfaces are nearby, the reflections will come in quick succession soon after the direct sound. Conversely, in a large room (e.g. gymnasium, concert hall) there is often a longer delay between the arrival of the direct sound and the first reflections. Also, the time between reflections may be longer. The (acoustic) absorptive characteristics of the surfaces within the room will determine the strength of the reflections relative to the direct sound and ultimately the reverberation time [31] of the room. A listener is sensitive to these various phenomena and uses them to derive an acoustic impression of the room [32].

There are two basic microphone parameters which can be manipulated to reduce the level of the camera noise recorded by a microphone: proximity and directivity. In the first, the relative proximity of the actor and the camera to the microphone is adjusted in order to minimize the level of the recorded camera noise. For example, a lavalier microphone can be placed in the actor's clothing. This helps to increase the signal-to-noise ratio of the recording. However, the timbre of the recording made in this manner is typically very unnatural.

Alternatively, a highly directional microphone can be used to focus-in on the voice of the actor while rejecting other sounds in the room (including the camera noise). The resulting recording will lack the reverberant and spatial information of the room and will have to be processed to try to simulate the lost information. Moreover in general one cannot expect to obtain more than 10 dB of broadband noise reduction using directional microphones in real rooms [33].

Both of the above methods are equivalent to reducing the amplitude of the acoustic path h_n relative to the paths h_{s_i} as shown in Figure 2.1. While these microphone techniques can help to substantially reduce the level of the camera noise, their effectiveness is limited and is offset by the conflicting needs of capturing room ambiance.

2.3 Acoustic Barriers and Blimps

An obvious solution to reducing camera noise in film soundtracks is to reduce the level of the mechanical noise produced by the camera. In fact, this has been done to a large extent over the past decades. However, the technological advances which have allowed for quieter camera operation have been offset by other advances which have reduced the noise and distortion in sound recordings. That is, although newer cameras may be quieter, the quality of film soundtracks has improved such that the camera noise is still audible. This is particularly true for digitally recorded soundtracks which have a nominal dynamic range of about 96 dB.

In the very early days of motion pictures with sound, camera noise was an extremely significant issue. To overcome the noise problem, cameras were often placed behind acoustic barriers or in acoustically isolated booths with a window through which the scene could be filmed [6]. Referring back to Figure 2.1, this is equivalent to reducing the path h_n. While this approach serves to reduce the camera noise in the soundtrack, it imposes serious limitations on the visual aspects of film. Basically, the approach requires

that the camera be in a fixed position throughout the scene. Ironically, the introduction of sound in motion pictures has been viewed by some as the main cause for the slow development of artistry in the visual aspect of films.

The modern version of the acoustic barrier is called a "blimp" which is an enclosure which encases the camera. Unlike the acoustically isolated booths described above, a blimp allows the camera to be mobile while filming a scene. However, any acoustic barrier requires mass in order to be effective [34] and so a blimp can significantly increase the size and weight of the camera, thus limiting its mobility. Moreover, blimps do not entirely eliminate the audibility of camera noise [5].

While camera manufacturers continue to reduce the level of noise produced by their cameras, this is understandably not their first priority, and they do not want to compromise image quality or mobility for lower noise. Also, as with many mechanical devices, wear of the parts over time may increase the noise produced by the camera. The use of a blimp implies additional costs at the time of filming and may not be a viable option if filming of a scene demands significant camera mobility. While reducing the noise at its source seems like a logical solution, it is not always an option.

2.4 Dolby 430 Series Background Noise Suppressor System

The Dolby 430 Series Background Noise Suppressor System is an analog signal processing device which is sometimes used to combat the effects of camera noise [4]. The Dolby device is intended to reduce the audibility of broadband noises such as wind or traffic rumble and was not designed with the goal of reducing camera noise [35]. Nonetheless, it is sometimes used for this purpose since no alternative signal processing approaches are readily available. The Dolby noise suppressor is based on the Dolby SR noise reduction system [36] which is used to reduce the background noise in analog tape recordings.

The Dolby system operates by dividing the input signal into two frequency bands each of which is followed by an expander circuit (i.e., a level dependent attenuator). In the lower frequency band the signal content below about 2 kHz is determined relative to the nominal signal level. If the level of the signal in this low frequency band is within ±10 dB of the nominal level, then nothing is done to the signal. However, if the level of the signal in the low frequency band is more than 10 dB below the nominal level, then a level dependent shelving filter [1] is applied to the signal. This low frequency shelving filter is flat for frequencies above 2 kHz, but can attenuate low frequency signals by as much as 18 dB. The depth of this shelving filter increases with decreasing low frequency input

signal level. A similar process is done simultaneously for the frequencies between 200 Hz and 8 kHz.

The effectiveness of this system relies on the fact that perceptually, a high level signal will mask a relatively low level noise occupying the same frequency range. Therefore, under these conditions there is no need for any noise reduction. However, as the signal-to-noise ratio decreases, the interfering noise will become more audible. To reduce the audibility of this noise, the input signal is filtered in proportion to its signal-to-noise ratio. The Dolby noise suppressor is most effective at reducing low level broadband noise in the short gaps that occur in speech signals. It does not however, provide any noise reduction outside these gaps, where there is a significant signal present. Also, for situations where the signal-to-noise ratio of the input signal is low, the Dolby system can create audible artifacts [35].

One important functional advantage of the Dolby system is that it does not require a reference signal (i.e., a separate recording of the interfering noise). This dramatically increases the potential usefulness of the system since there is no need for an additional recording of the noise source at the time of filming. Also, given this, the system could in theory be used to reduce noise in the restoration of old films. Another feature of the Dolby system is that it allows the user to directly control the amount of processing applied to the input signal. However, it does not have the ability to be self-adaptive to changes in the camera noise.

2.5 Automatic Dialog Replacement - Dubbing

There are situations where, despite the use of various microphone techniques, acoustic barriers, and post-processing, the level of the camera noise in a film soundtrack is still deemed to be unacceptable. In this situation the only recourse is dubbing or automatic dialog replacement (ADR). ADR is the procedure whereby the actors' dialogue is re-recorded in a quiet environment after the filming is completed [1].

Typically, ADR[†] is done in a dubbing theatre where the actors recite their dialogue while watching a projection of their previously filmed performance. Clearly this process completely eliminates any problem of camera noise. However, ADR can produce several undesirable side effects. In the ADR process the actors must carefully match their re-

[†] The "automatic" component of ADR appears to be the automatic synchronization of the newly recorded dialogue with the image.

recorded dialogue to the image of them talking. Any mismatch between the sound of their voices and the image of their lips moving might be noticeable and annoying to the audience. Also, dubbing the dialog for a given scene in a film requires that the actors recreate the emotional setting which was present at the time of filming. Although it may solve a technical problem, ADR can compromise the actor's performance.

While ADR resolves one technical problem for the sound recording engineer, it does introduce another. One of the goals when recording dialogue for a film is to also capture the ambiance of the room (i.e. background noises, reverberant characteristics, relative locations of the actors, etc.). Since ADR is done in a sound recording studio, rather than at the location where filming occurred, the engineer is faced with the task of trying to recreate the ambiance of the room. Done incorrectly, this results in a sudden change in the *character* of the soundtrack which occurs only for those portions of the dialogue which have been dubbed. Again, if this is noticeable to the audience, the illusion of reality created by the film may be destroyed.

A final but not unimportant consideration in ADR is that of cost. The combination of salaries (especially those of the actors and director) and the rental of the necessary facilities implies that ADR can be rather costly. The cost of ADR for a typical film is on the order of US$50,000 [37]. With hundreds of films (and television programs) being made each year, it is evident that millions of dollars are spent annually on ADR. Furthermore, the time required for dubbing may delay the overall production of the film. As such, a more cost-effective and less time consuming solution is desirable.

2.6 Other Signal Processing Attempts at Reducing Camera Noise

In this thesis, various signal processing techniques are examined for reducing camera noise. In this section we briefly describe other signal processing methods which have been previously proposed for reducing camera noise. Each of these methods was found to be ineffective and was subsequently abandoned.

2.6.1 Attempts by SAIC

Several years ago (circa 1995), the IMAX corporation funded a research effort to study ways of reducing camera noise [5]. The research was conducted by the Science Applications International Corporation (SAIC) in the U.S.A. SAIC explored the possibility of using active noise cancellation to reduce the acoustic noise output of the IMAX-3D camera. The system effectively consisted of an "electronic blimp". Unfortunately, the

method did not provide a significant amount of noise reduction, and so work on this approach was halted. The findings of Chapters 3 and 4 of this thesis suggest that a probable cause of the poor performance was the distributed nature of the camera noise.

A second method was also examined by SAIC wherein they applied signal processing techniques directly to the film soundtrack. This is in keeping with the approach proposed in this thesis. SAIC proposed the use of adaptive noise cancellation to reduce the level of the camera noise. It was found that this method did not provide sufficient noise reduction and so it was rejected. This same approach is examined in this thesis (see Chapter 4) and is shown to be unsuccessful due to the low inter-channel coherence resulting from the distributed nature of the camera noise. It should be noted that the author was unaware of the work done at SAIC until after the work described in Chapter 4 was completed.

A final method explored by SAIC involved a noise reduction scheme based on the use of neural networks. Like the methods proposed in this thesis, the scheme was a single input approach, thus making it more suitable for the camera noise application. This method was also abandoned because it did not provide a useful amount of noise reduction and it severely distorted the desired signal.

Unfortunately, details regarding the approaches developed by SAIC are not available since all of the results are contained in proprietary reports which are not publicly available. The collaboration between SAIC and IMAX was terminated with the conclusion that signal processing techniques examined by SAIC for reducing camera noise in film soundtracks are not viable.

2.6.2 Commercially Available Broadband Noise Reduction Systems

There are several commercially available noise reduction systems for restoring gramophone recordings. These systems perform several tasks including; click removal, correction of pitch errors, removal of low frequency noise pulses due to breakages in the surface of the disc, and broadband noise reduction [20,30,38,39,40,41,42]. These systems can be very effective at removing noise from gramophone recordings. The broadband noise removal components of these systems are generally based on spectral subtraction (see Chapter 5), but specific details of their operation are proprietary. These systems have been under development for many years and it is interesting to ask how these more mature technologies might perform at removing camera noise. Unfortunately, these noise reduction systems are very expensive, and the author did not have direct access to them. However, staff at two film studios have experimented with these systems and subse-

quently concluded that they were not suitable for the task of removing camera noise from film soundtracks [4,43]. In fairness to these noise reduction systems however, they were not designed for this application and do not take advantage of the repetitive nature of camera noise.

2.7 Summary

In this chapter it was seen that some options for reducing or eliminating camera noise currently exist. However, these solutions may compromise certain artistic aspects of the film, create other technical problems, or incur significant costs. Previous attempts at reducing camera noise have been unsuccessful, and systems designed for restoring gramophone recordings are not effective for the camera noise application.

3. CHARACTERIZATION OF CAMERA NOISE

3.1 Introduction

Before exploring potential signal processing approaches for reducing the audibility of camera noise in film soundtracks, it is useful to characterize the underlying properties of the noise. To this end, a series of comprehensive acoustic measurements of a professional film camera were made at the studios of the National Film Board of Canada (NFB) in Montreal. The results of the measurements provide valuable information regarding the fundamental properties inherent to camera noise, as well as a database of recordings for developing and evaluating noise reduction schemes.

Prior to making the acoustic measurements, discussions were held with staff members of the NFB [3,4] from which it was determined that the following parameters were likely to have an effect on the level and quality[‡] of the camera noise: make of camera, film size (i.e. 16 mm, 35 mm, etc.), type of lens, film stock, and location within a reel of film. Therefore, measurements were conducted to examine and evaluate any effect on the camera noise due to these parameters.

Unfortunately, at the time the measurements were made, only one make of camera was available at the NFB and therefore, differences in camera noise due to the make of the camera or the film size could not be evaluated. It was believed however, that the results of these measurements would be directly applicable to other makes and models of cameras.

Since the time that the measurements were made at the NFB, the IMAX™ corporation provided the author with recordings of the noise produced by several of their cameras. As well, IMAX allowed the author access to their 3-D camera in order to make measurements and recordings. Due to various constraints, measurements of the IMAX cameras were not as comprehensive as those conducted at the NFB. Nonetheless, the results of the IMAX measurements support the assumption that the results of the NFB measurements are applicable to other cameras.

[‡] Here the term *quality* refers to the characteristics of the noise and not to any parameter which would influence one's preference or dislike for the camera noise.

3.2 Description of the NFB Measurement Set-Up

The measurements of the camera noise were conducted in a large mix-down studio at the NFB in Montreal. The mix-down studio, which is used to mix the final soundtrack of a film, consisted of a very large room with high ceilings and a projection screen at one end of the room. The studio was acoustically treated to reduce the level of background noise and to minimize the reverberation within the room. Ideally, the measurements would have been conducted in an anechoic chamber so that only the sound emanating directly from the camera would be measured. Non-anechoic conditions imply that acoustic reflections from nearby surfaces will inevitably be included in the measurements. Although measurements in an anechoic chamber were not possible, the effects of any acoustic reflections were lessened by the large size of the room which made it possible to place the camera and measurement microphones such that the nearest surface (other than the floor of the room) was approximately 7 m away. Therefore, because of the large distance between the microphones and the reflecting surfaces, and because these surfaces were acoustically treated, the reflected acoustic energy was greatly attenuated at the microphone. As such, it was felt that the studio provided very acceptable conditions for the acoustic measurements described here.

The camera which was analyzed was an AATON - Regular 16 mm camera with an AATON M3908 MAG-A housing the film. This camera operates at the typical film rate of 24 frames per second. Eastman EXR 7245 color negative film was used in all of the measurements of camera noise.

Two Brüel and Kjær Type 4165 measurement grade microphones were used to measure the noise from the camera. The outputs of the microphones were connected to two microphone preamplifiers before going to a Sony PCM-7030 DAT recorder. The sampling rate, f_s, of the DAT recorder was set to 48 kHz for all of the measurements. This is the sampling rate which is typically used in the film industry since it is an integer multiple of the rate (24 frames/sec) at which the film in the camera operates. It will be seen later in the thesis that this relationship between the sampling rate of the audio recordings and the frame rate of the camera provides some useful benefits in the noise reduction process.

Figure 3.1 provides an overview of the measurement setup. As can be seen from the figure the two microphones were placed at 90° to each other. This was done in order to measure the directivity of the camera noise as well as the distributed nature of the camera noise. The front of the camera (i.e., where the lens is pointing) was designated to be 0°, thus making the rear of the camera 180°. The left side of the camera was designated to be

90° while the right side of the camera was 270°. Each microphone was located at a distance of 1 m from the centre of the camera.

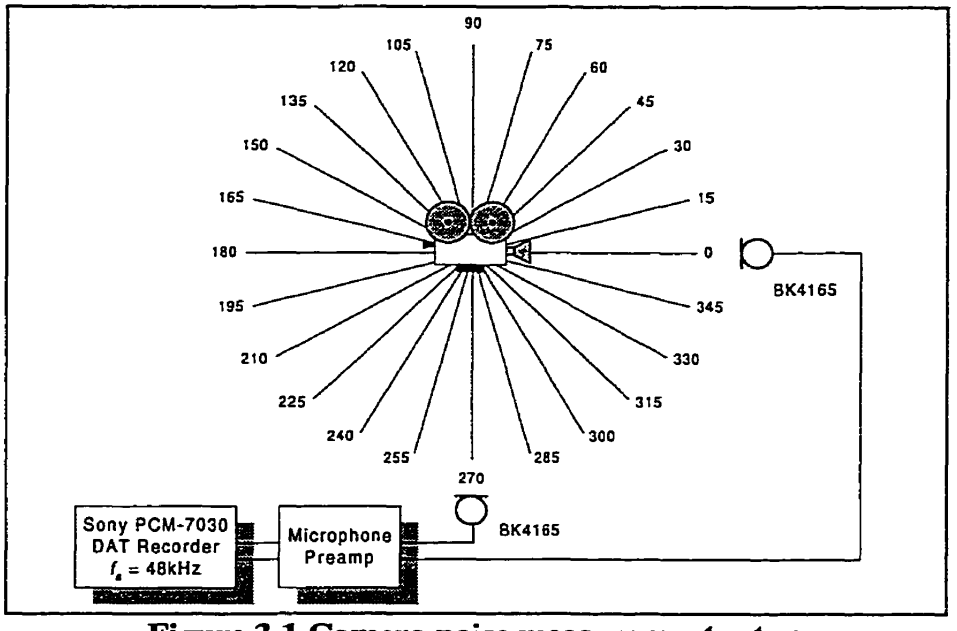


Figure 3.1 Camera noise measurement setup

3.3 Calibration of Microphones

Prior to making any measurements, the two microphones as well as the two channels of the measurement system were calibrated using a Brüel and Kjær Type 4230 calibrator. The calibrator provides a 1 kHz acoustic sinusoidal signal of precisely 94 dBSPL (re: 20 μpascal) thus making it possible to measure the absolute level of the camera noise. By adjusting the input gain on the DAT recorder, full-scale (16 bits PCM) on the recorder was set to correspond to 94 dBSPL. Using this as a reference, the absolute sound pressure level of all subsequent recordings could be determined.

It was assumed beforehand (and later confirmed) that the level of the noise emitted by the camera would be well below 94 dBSPL and therefore using this level as full-scale would not maximize the potential dynamic range of the measurement system. Therefore, during the calibration process, a 20 dB attenuator was inserted into the signal path between the preamplifier and the DAT recorder. During the subsequent measurements, the 20 dB attenuator was removed. By inserting the attenuator into the signal path during the calibration process, the sound pressure level corresponding to full-scale on the DAT recorder was reduced from 94 dBSPL to 74 dBSPL. The result was an effective 20 dB increase in the dynamic range of the measurement system thus ensuring that the subsequent

acoustic measurements would not be limited by the noise floor of the measurement system. The calibration process was done separately for each microphone.

3.4 Description of NFB Measurements

As mentioned in Section 3.1, several variables were believed to contribute to the level and quality of the camera noise. However, given that only one camera was available, only a subset of these variables could be examined. A series of measurements was conducted to investigate these parameters.

3.4.1 Background Noise of Measurement System

As a first measurement, the level of the background noise of the measurement system was determined. Here the *system* consisted of the room as well as all of the electronic components (microphones, pre-amps, DAT recorder) used in the recordings. For this measurement, the camera was turned off and a 30 s recording was made of the ambient noise. This recording therefore included the noise due to the room as well as the noise floors of the various electronic components in the recording chain. The spectrum of the measured background noise is shown in Figure 3.2 with the vertical axis indicating the sound pressure level (re: 20 μpascal). The figure shows the level of the noise decreasing steadily for increasing frequencies. This is typical of the acoustic background noise found in this type of room [34].

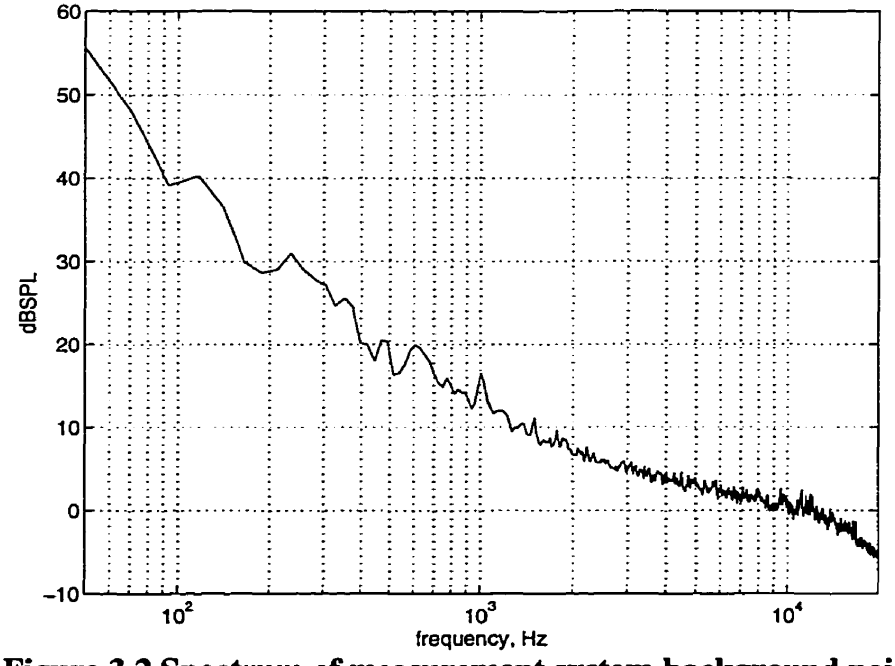


Figure 3.2 Spectrum of measurement system background noise.

The background noise of a system is frequently expressed in terms of the overall A-weighted sound pressure level which was 34 dBA in this case. The background noise measurement of Figure 3.2 serves as baseline by which it may be determined whether a spectral component in a given measurement is due to the camera or the measurement system.

3.5 Typical Camera Noise

3.5.1 Basic Characteristics of Camera Noise

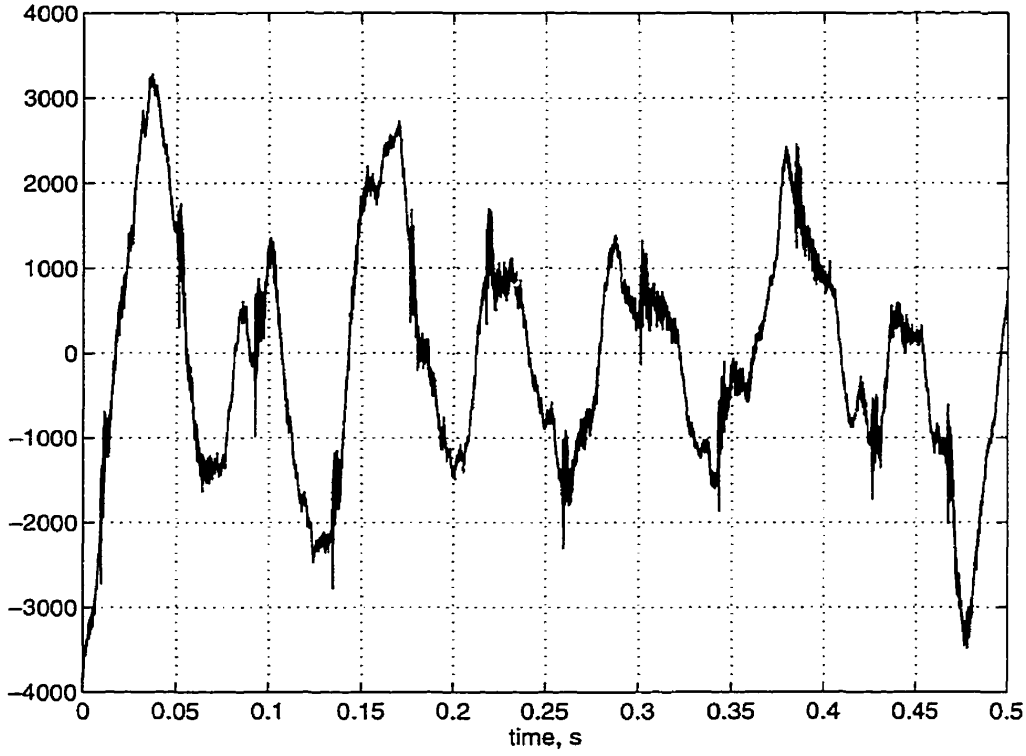


Figure 3.3 Time waveform of camera noise.

In this section the results of the acoustic measurements are presented in order to provide a general overview of some of the main characteristics of the camera noise. All of the results shown in this section pertain to measurements of the camera noise taken at 0° with the zoom lens mounted on the camera. Figure 3.3 shows an example of the time waveform of typical camera noise. It should be noted that the waveform represents only the camera noise. That is, there is no desired signal, such as speech in the waveform. However, there is a significant amount (relative to the level of the camera noise) of low frequency room noise in the waveform. Therefore, the signal depicted in Figure 3.3 was

highpass filtered to reduce the level of the room noise and allow a better examination of the nature of the underlying camera noise. The resulting waveform is shown in Figure 3.4.

The highpass filter consisted of a 100 tap FIR (finite impulse response) filter with a cut-off frequency of 70 Hz. The filter was designed to be linear phase thereby keeping the time domain waveform intact. That is, there is no "smearing" of the waveform due to the phase response of the filter which is an important consideration given the repetitive nature of the camera noise.

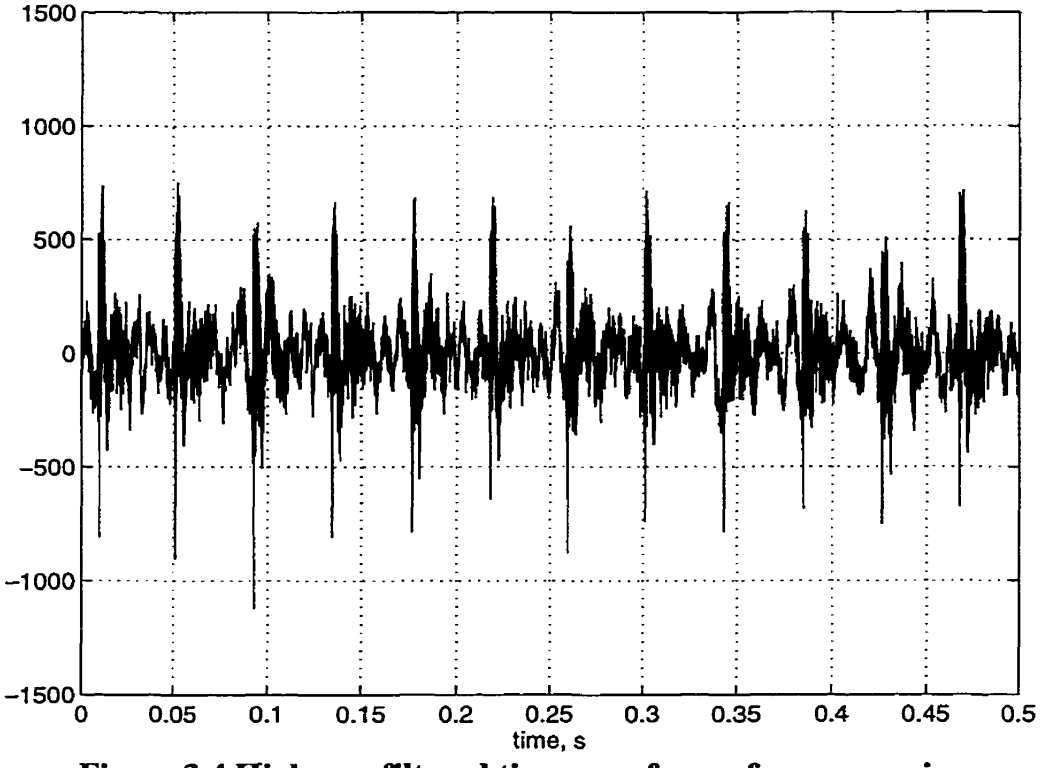


Figure 3.4 Highpass filtered time waveform of camera noise.

As can be seen, the camera noise consists of a series of regularly spaced pulses. The time between the pulses is 1/24th of a second which corresponds to the film rate of the camera (i.e., 24 frames per second). Since the audio recordings were made at a sampling rate of 48 kHz the time between pulses is 2000 samples. This fact will be used to advantage in the noise reduction schemes described later. While the individual pulses are similar in their appearance, it is clear that each pulse is unique. Also, while each of the individual pulses appears to contain a large amount of the noise power, a significant amount of noise is also present between the pulses.

Figure 3.5 provides a close-up view of 5 pulses. From the figure one can more readily see the differences between individual pulses. Also, the portion of the camera noise which lies between the peaks of the pulses is seen more clearly in the figure. While there are some similarities in the structure of the sections of noise between the pulses, it can be seen that, again each of these sections is unique. The uniqueness of the individual pulses and the noise energy between the pulses will play an important role in determining the effectiveness of the noise reduction scheme described in Chapter 4.

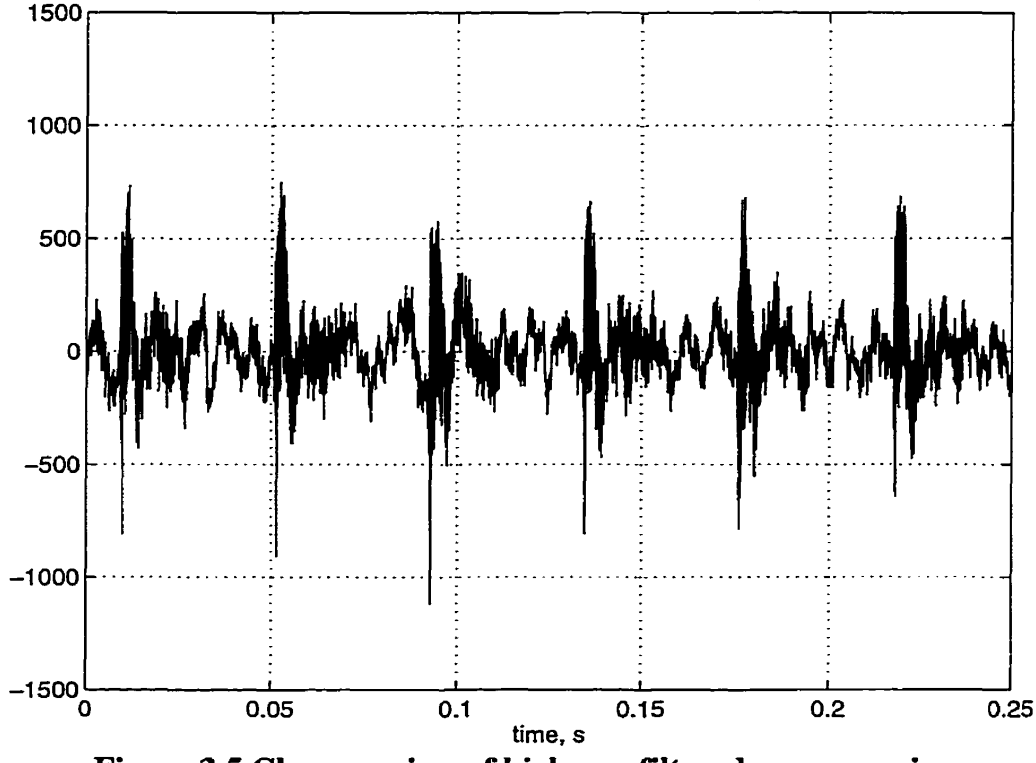


Figure 3.5 Close-up view of highpass filtered camera noise.

Figure 3.6 provides a close-up view of a single pulse. Interestingly, from this view, it is now clear that the "pulse" of the camera noise is actually made of several peaks. Furthermore, the section of noise following the peak of the pulse appears to be quite random in nature.

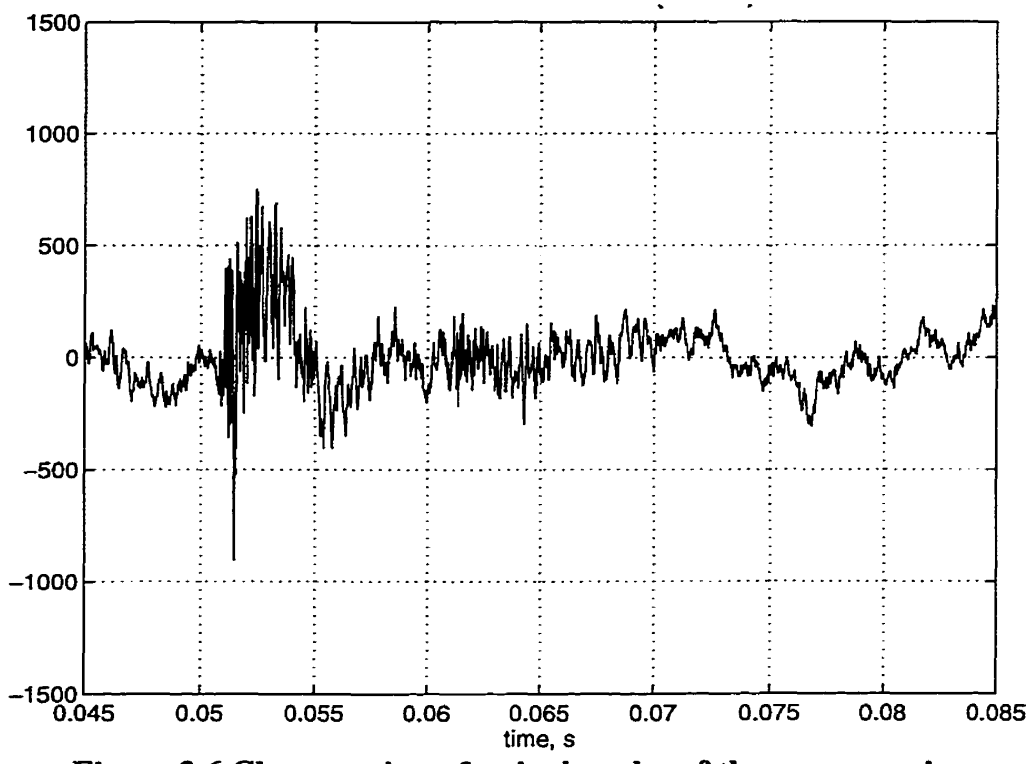


Figure 3.6 Close-up view of a single pulse of the camera noise.

So far we have seen only time domain plots of typical camera noise waveforms. Figure 3.7 provides a view of the power spectrum of typical camera noise. The power spectrum was obtained by averaging over 170 pulses (about 7 s) of camera noise measured at 0°. The power spectrum was derived using Welch's method [44] based on modified periodograms with a Hanning window and 50% overlapping of the time segments. The upper curve in the figure corresponds to the camera noise, while the lower curve is the system background noise described earlier. It can be seen that the camera noise has a broadband spectrum, although the level of the camera noise becomes insignificant (relative to the background noise) below about 100 Hz.

The power spectrum of Figure 3.7 does not reveal any obvious harmonic structure to the camera noise which would be easy to detect and remove. While there appears to be a peak in the spectrum at about 3800 Hz, this is not consistent and is merely particular to this measurement.

Given that the camera operates at a rate of 24 frames per second, one might reasonable expect to find spectral lines in the power spectrum related to this rate. These spectral lines do occur, but only if the power spectrum is measured over many pulses, rather than over single pulses as was done in Figure 3.7. This matter is discussed in greater detail in Appendix A.

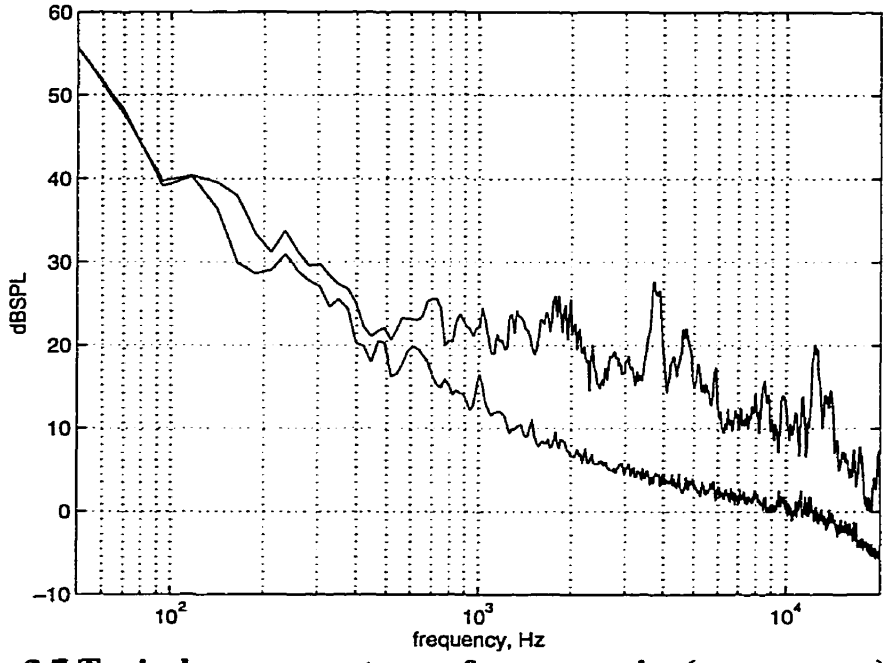


Figure 3.7 Typical power spectrum of camera noise (upper curve); system background noise (lower curve).

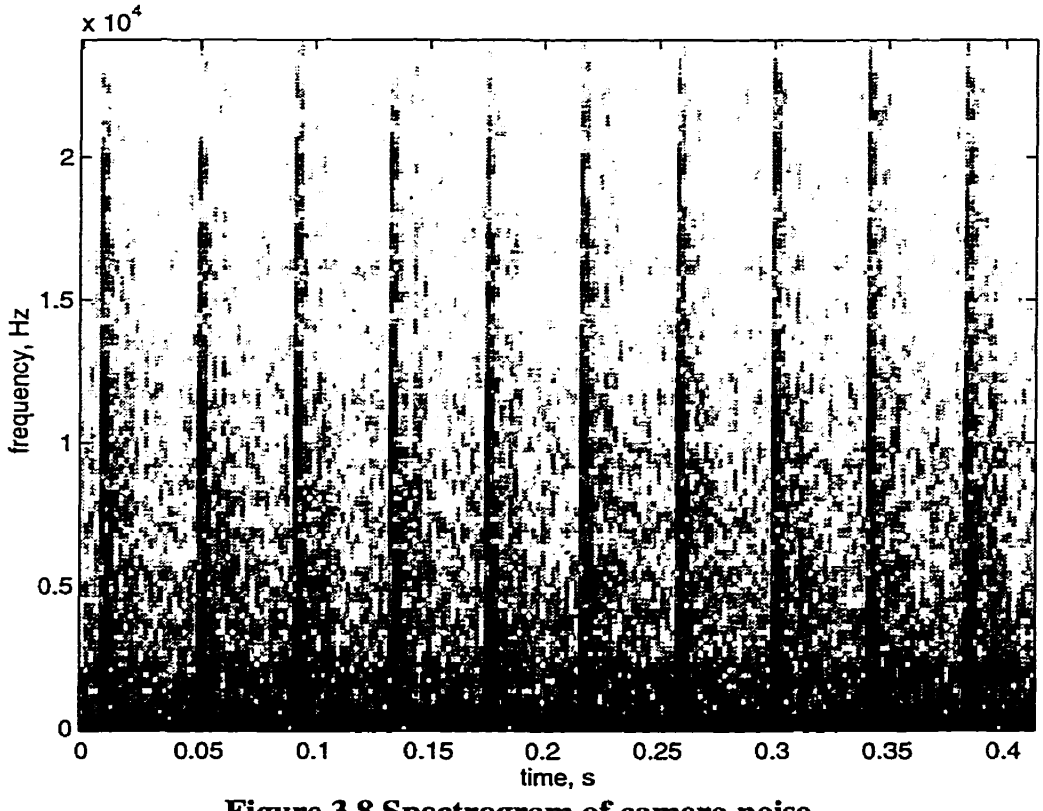


Figure 3.8 Spectrogram of camera noise.

An alternative way to view the camera noise is in the form of a spectrogram. Figure 3.8 shows the spectrogram over several camera pulses (a Hanning window was used with 256 point FFT's). In this plot the horizontal axis represents time while the vertical axis represents frequency. The log amplitude of the signal is depicted by the shading of the figure. Lighter shades of gray represent lower amplitudes while darker shades represent higher amplitudes, with black being the highest amplitude.

In the figure, it is easy to see where the peak energy of the pulses occur. The peaks create a broadband noise which extends over the entire audible spectrum. Following each peak is a region over which the level of the noise rapidly decays, while prior to each peak, the camera noise is low. This is particularly true in the higher frequencies. This structure of the camera noise will be exploited in some of the noise reduction algorithms to be described in the sequel.

3.5.2 Directivity of the Camera Noise

The purpose of the next set of measurements was to determine the directivity of the camera noise. The directivity of the camera noise is important since it directly affects the amount of noise picked up by the microphone recording the actor's dialogue. If the camera noise is not omni-directional, then the noise picked up by the microphone will vary as the relative positions of the camera and the microphone change. That is, there may be directions for which the camera noise is stronger and so if the microphone is placed in that position, the level of the recorded noise will be higher than for other positions. Also, the directivity of the camera noise is particularly important when considering a noise reduction scheme such as LMS (least-mean-square) based adaptive noise cancellation which requires that a recording of the interfering signal (i.e., the camera noise is this case) be available with little or none of the desired signal mixed with it. Therefore, when deciding where to place the transducer for this recording, it is useful to know where the camera noise is loudest.

With the camera operating, the resulting noise was measured at 15° intervals using the setup depicted in Figure 3.1. By using two microphones, the time required to do these measurements was cut in half thus reducing the effect of any possible time varying parameters. For these measurements the camera was fitted with the zoom lens and the height of the microphones was set to 51 inches which was equal to the height of the centre of the camera lens. Recordings of 10 s duration were made at 24 angles in the horizontal plane. The recordings were filtered on an octave-band basis and polar plots de-

picting directivity were generated for the octave bands extending from 125 Hz to 8000 Hz.

Figure 3.9 shows the directivity of the camera noise averaged over the octave band centered at 125 Hz. Again, 0° represents the front of the camera which is the direction where the lens is pointing. In this plot (and for all subsequent directivity plots) the shaded area represents the level of the noise created by the camera. For a given angle, the distance from the centre of the plot to the edge of the shaded area indicates the level of the noise in decibels. For example, referring to Figure 3.9, the level of the noise at 0° is about 42 dB. As one might expect given the relatively small size of the camera, the noise is nearly omni-directional in the 125 Hz octave band.

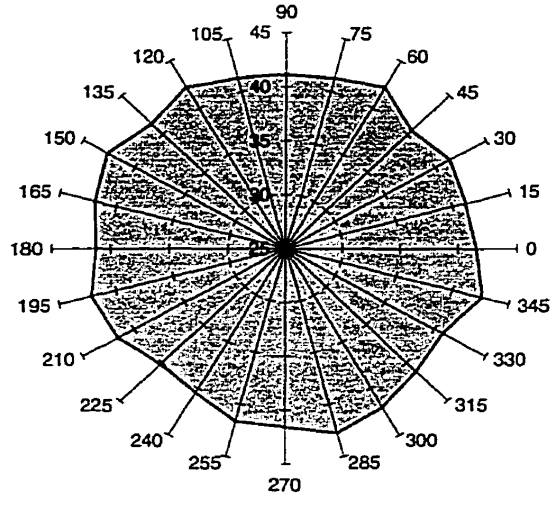


Figure 3.9 Directivity at 125 Hz

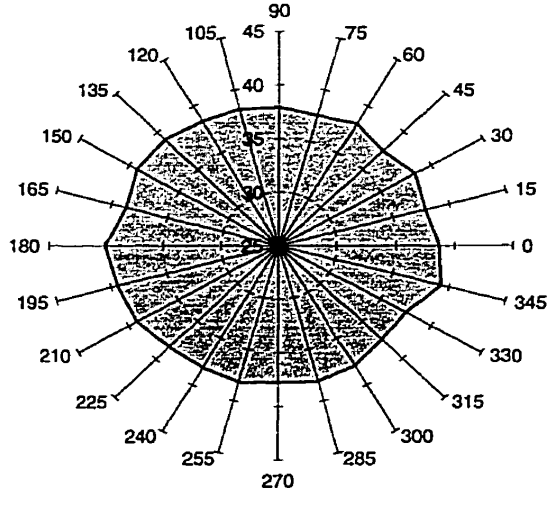


Figure 3.10 Directivity at 250 Hz

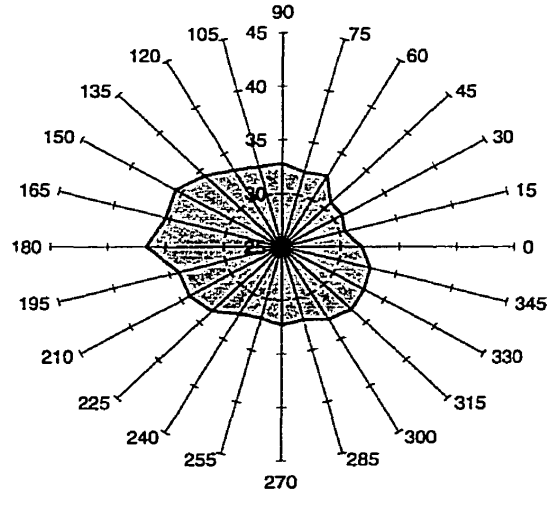


Figure 3.11 Directivity at 500 Hz

Figure 3.10 shows the directivity of the camera noise in the 250 Hz octave band. Again the noise is omni-directional, although the level is somewhat attenuated relative to the 125 Hz octave band.

The directivity of the camera noise in the 500 Hz octave band is given in Figure 3.11. The level of the noise in this octave is significantly lower than for the two previous octave bands. Also, the noise has become somewhat directional, with higher levels of noise found towards the rear of the camera.

Figure 3.12 shows the directivity of the noise in the 1000 Hz octave band. In a manner similar to the 500 Hz octave band, the noise in the 1000 Hz octave band is somewhat directional, with the higher noise levels at the rear of the camera. Interestingly, the overall level of the noise has increased in this octave band relative to the 500 Hz octave band.

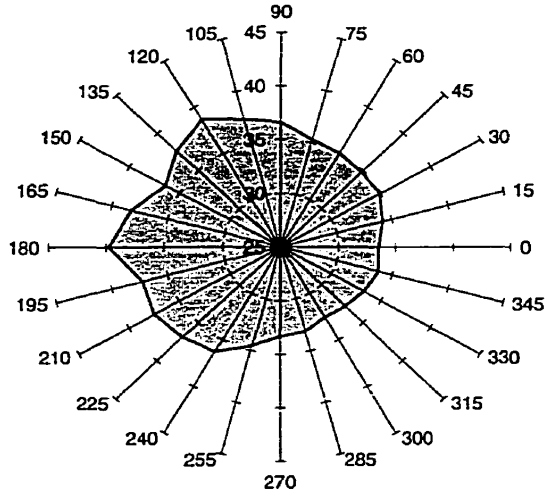


Figure 3.12 Directivity at 1000 Hz

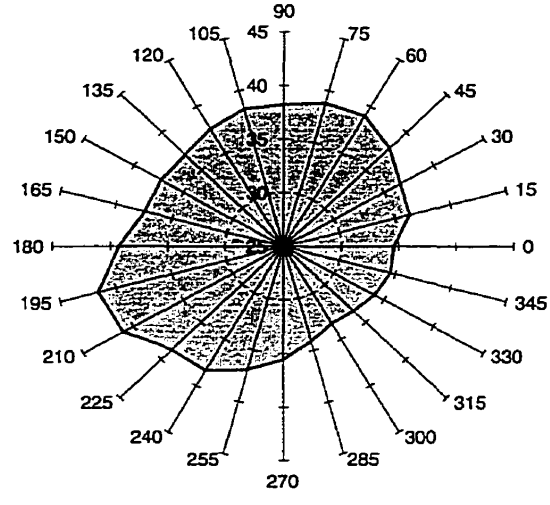


Figure 3.13 Directivity at 2000 Hz

The noise in the 2000 Hz octave band shows a very different directivity pattern as compared to any of the other octave bands. From Figure 3.13 it can be seen that the level of the noise is greater towards the front-left and the rear-right of the camera. The level of the noise is this octave band is greater than at 1000 Hz and is comparable to the level at 4000 Hz.

The directivity patterns of the noise at 4000 Hz and 8000 Hz are similar. For both octave bands the noise is greater towards the rear and is particularly strong on the left side of the camera Figure 3.14 shows the directivity at 4000 Hz while the directivity at 8000 Hz is shown in Figure 3.15.

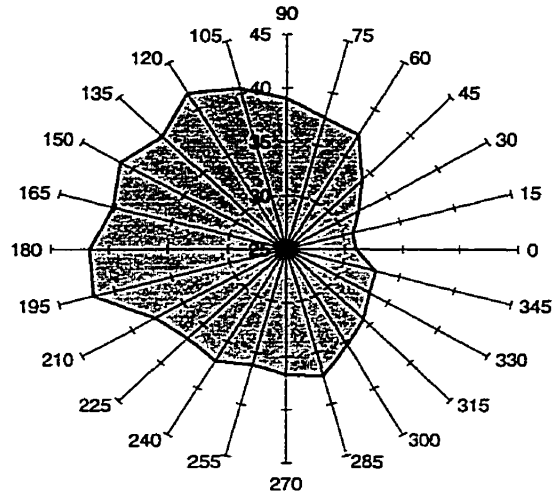


Figure 3.14 Directivity at 4000 Hz

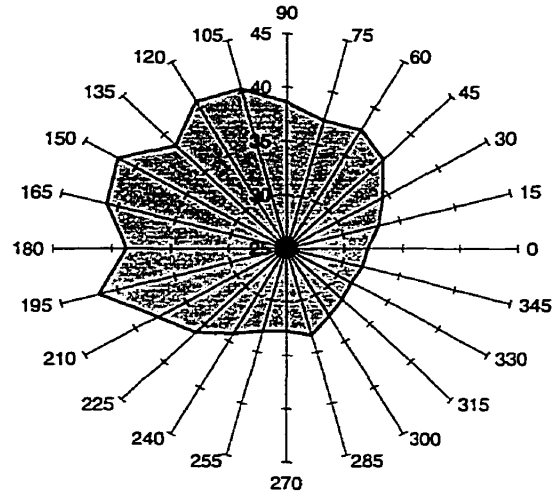


Figure 3.15 Directivity at 8000 Hz

From the polar plots shown in the above figures, it can be concluded that the directivity of the camera noise varies significantly with frequency. At some frequencies, differences of as much as 10 dB can be found for different angles as the relative positions of the camera and microphones change. This implies that any successful noise reduction algorithm must be capable of adapting to the resulting changes in the level and spectrum of the camera noise. Furthermore, the results of the measurements described in this section indicate that there does not appear to be any single best angle at which to place a microphone in order to record the reference signal for an adaptive noise cancellation approach to reducing the noise. More importantly, the relatively high directivity of the camera noise at some frequencies suggests that the camera noise is not behaving as a point source, but rather as a distributed source. This matter is further investigated later in this chapter.

3.5.3 Effect of Camera Lens

It was believed that the type of lens used during filming could have an effect on the resulting noise emitted by the camera [3,4]. Therefore, measurements of camera noise were made with two types of lenses mounted on the camera: a prime lens and a zoom lens. The measurements were again made at 15° intervals around the camera as described in Section 3.2, and they were made on a single reel of film so that only the effect of the lens would be measured. That is, it was assumed beforehand that different film stocks could have an effect on the camera noise (see Section 3.5.4 *Effect of Film Stock*).

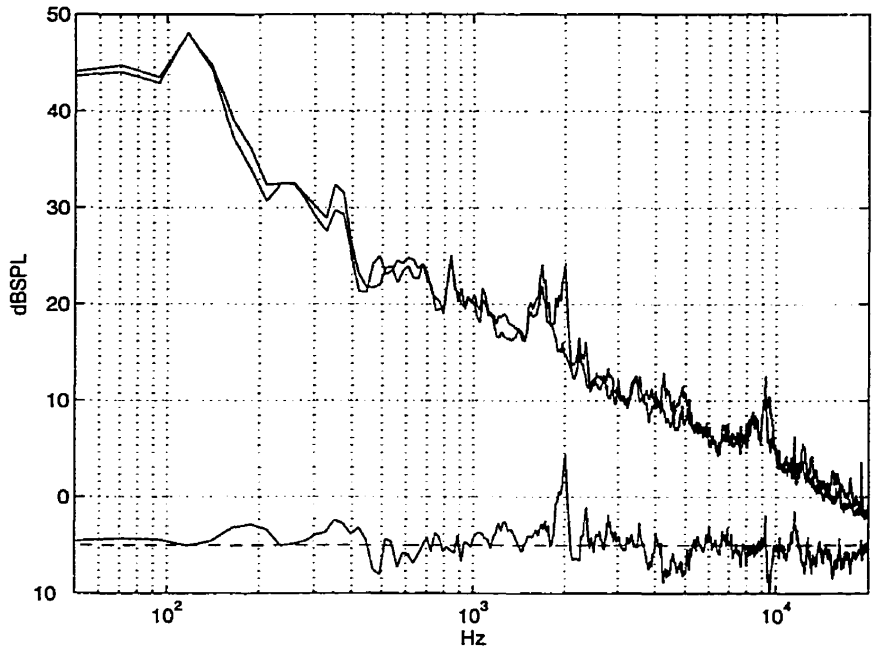


Figure 3.16 Effect of lens type on the spectrum of the camera noise.

Figure 3.16 shows a typical example of the camera noise with a prime lens versus a zoom lens. It can be seen that there is no apparent change in the overall sound pressure level of the noise as a result of changing lenses. There are however, differences in the fine structure of the power spectrum. The lower plot (centered on the -5 dB grid line for clarity of presentation) shows the difference between the two power spectra. Although the differences in the noise spectra are relatively small for the most part, there are points in the spectra which differ by as much as ±5 dB. These differences are large enough to be audible and so a noise reduction method that can adapt to these changes appears to be warranted. It is possible that the differences seen in Figure 3.16 may be due to the location in the film rather than due to the lens being used. This matter will be examined in Section 3.5.5.

3.5.4 Effect of Film Stock

Discussions with camera experts at the NFB revealed that significant differences in the camera noise could be expected with changes in the film stock used in the camera. Therefore, complete directivity measurements (as described in Section 3.5.2) were made for two different reels of film. Both reels consisted of the same type of film (Eastman EXR 7245 Color Negative Film).

The power spectra for the two film stocks are shown in Figure 3.17 for an angle of 0°. It is apparent that significant changes can occur to the spectrum of the camera noise as a

result of changing the film stock. Both the overall sound pressure level of the noise and the structure of the spectrum have changed significantly. It should be noted that the figure represents the difference in the camera noise at only one angle. Much larger differences can occur at other angles as can be seen in Figure 3.18 which shows differences in camera noise at six different angles.

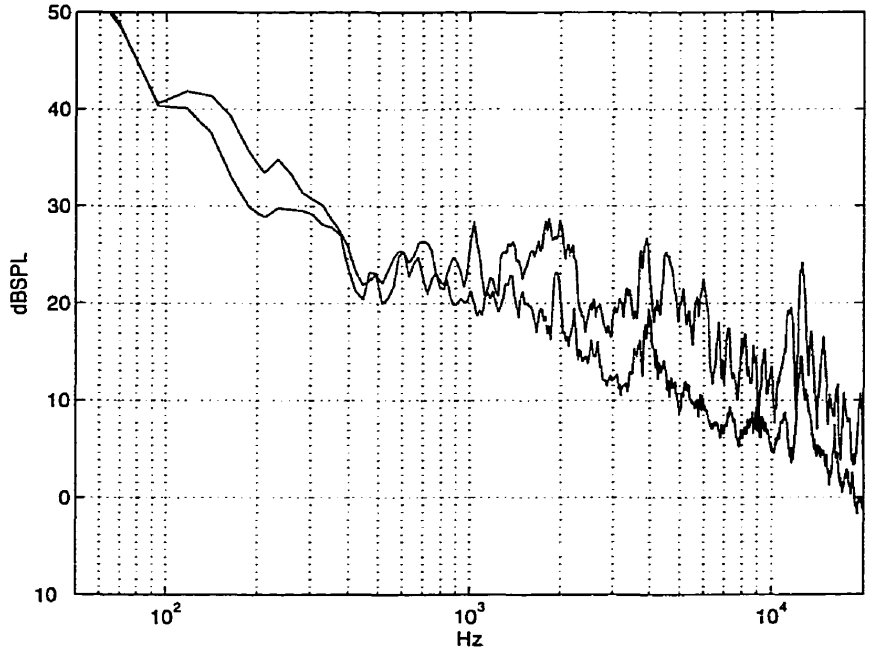


Figure 3.17 Effect of film stock on the spectrum of the camera noise.

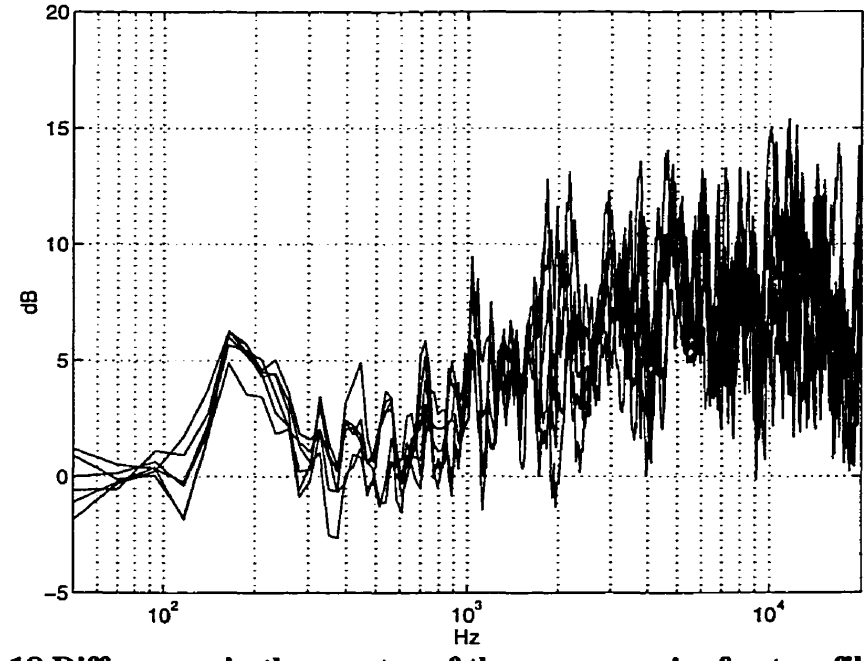


Figure 3.18 Differences in the spectra of the camera noise for two film stocks at different angles.

It can be seen that differences in the noise spectra of as much as 15 dB were measured as a result of changing film stocks. The changes to the power spectra are relatively small for frequencies below about 300 Hz. Above this point however, the changes tend to increase with increasing frequency. It is important to note that the change in the spectrum of the camera noise is different for each angle. That is, the directivity of the noise has also changed as a result of changing the film stock. Also, it should be stated that it may be possible for even greater differences to occur for different cameras, different types of film, or other reels of the same type of film. These points were not explored in these measurements.

Figure 3.17 and Figure 3.18 clearly demonstrate that changing film stocks can have a significant effect on the resulting camera noise. Therefore, any method developed for reducing camera noise in film soundtracks must be able to adapt to these changes.

3.5.5 Effect of Location Within a Reel of Film

The possibility that the power spectrum of the camera noise might change due to the location within the reel of film was also investigated. To do this, a recording of the camera noise was made over the length of an entire 122 m reel of film (about 5 min). For this recording the microphone was placed at 0° and the zoom lens was mounted on the camera. This recording was then analyzed to examine the degree to which the camera noise changes over short intervals of time, as well as for longer periods.

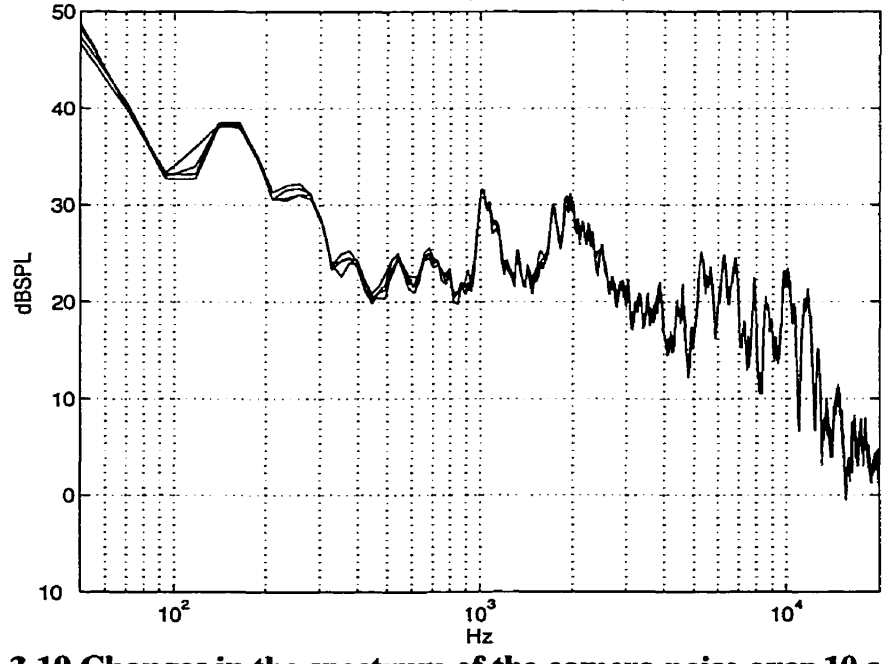


Figure 3.19 Changes in the spectrum of the camera noise over 10 s intervals.

Figure 3.19 shows four different power spectra of the camera noise measured for the same reel of film. The power spectra were measured at 10 s intervals and were averaged over 2 s. These power spectra were again derived using Welch's method [44] using a 2000 point Hanning window with 50% overlap. It can be seen that the power spectrum changes very little over this period (about 32 s) of time. Therefore, it appears that for a given set-up (camera, lens, film stock, etc.), the power spectrum of the camera noise is relatively constant over short periods of time.

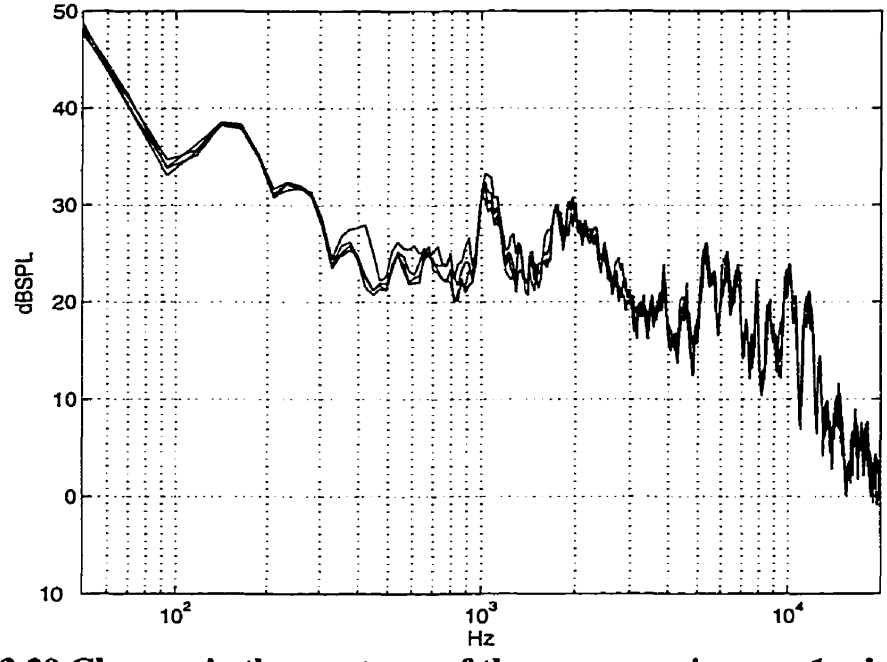


Figure 3.20 Changes in the spectrum of the camera noise over 1 min intervals.

Figure 3.20 is similar to Figure 3.19 except that the power spectra were measured over 1 minute intervals rather than 10 s intervals. Therefore, in this case the four measurements were taken over a total period of about 3 minutes. Again, the variations in the power spectrum of the camera noise are rather small over this time period. As might be expected however, the variations are somewhat larger over these longer time intervals.

The measurements described in this section reveal that, for a given reel of film, the power spectrum of the camera noise does not change very much over time. This feature of the camera noise will be exploited in the noise reduction method based on spectral magnitude estimation described in Chapter 5.

3.5.6 Recordings of Dialogue

At the time that the audio recordings were made to characterize the NFB camera noise, additional recordings were made to capture examples of dialogue corrupted by camera

noise. These recordings of speech in the presence of camera noise were intended to be used as test sequences to evaluate the various noise reduction algorithms. During the recordings the talker stood directly in front of the camera at a distance that was deemed by the camera operator to be typical for a *close-up*. This corresponded to a distance of about 1 meter from the camera to the talker. The microphone was placed as close as possible to the talker, but not so close that it would be visible in the camera's field of view. Therefore the recordings represented a realistic configuration for filming a close-up shot in a film. In addition, a second microphone was placed nearer to the camera in order to record a "worst case" condition of dialogue corrupted by camera noise.

Twenty test sentences from the Revised Harvard Test Sequences [45] were used as dialogue for these recordings (see Appendix B). The Harvard Sentences are commonly used in subjective tests to determine speech intelligibility or to evaluate speech quality.

The same 20 sentences were also recorded in the absence of any camera noise. These "noise free" recordings provided a reference of uncorrupted speech to be used in the evaluation of the various noise reduction schemes.

From the recordings of speech corrupted by camera noise it is possible to estimate typical and worst case signal-to-noise ratios for the dialogue. To do this, the recordings were first highpass filtered at 70 Hz using an FIR filter. This was done to remove the room noise that was earlier seen to dominate the recordings at frequencies below about 100 Hz. The levels of the speech signal and the camera noise were then determined from the highpass filtered waveforms. The broadband signal-to-noise ratio for the "typical" close-up recording was found to be 37 dB while the "worst case" signal-to-noise ratio was 22 dB. The value of a "typical" signal-to-noise ratio was derived from the recording using the microphone nearer to the talker, while the "worst case" value was measured from the recording using the microphone nearer to the camera. It is important to note that these represent *average* signal-to-noise ratios in that the level of the speech (dialogue) is expected to vary significantly within an actor's performance and therefore, parts of the soundtrack may have a significantly poorer signal-to-noise ratio. Furthermore, it was seen earlier that the film stock has a significant effect on the level of the camera noise. Therefore, it is expected that the above values of signal-to-noise ratios do not represent the worst case.

The signal-to-noise ratio values reported above do not take into account the frequency content of the speech and the camera noise. Figure 3.21 shows the average power spectra signal-to-noise ratios versus frequency as measured at the two microphones. The upper

curve represents the "typical" signal-to-noise ratio versus frequency, while the lower curve is for the "worst case" signal-to-noise ratio. It can be seen from the figure that the values for the signal-to-noise ratios (37 dB and 22 dB) quoted earlier are dominated by the 100 Hz to 2 kHz frequency range. That is, the signal-to-noise ratios in this range are much greater than for higher frequencies. At higher frequencies, the average "typical" signal-to-noise ratio is about 20 dB while the "worst case" signal-to-noise ratio is about 5 dB.

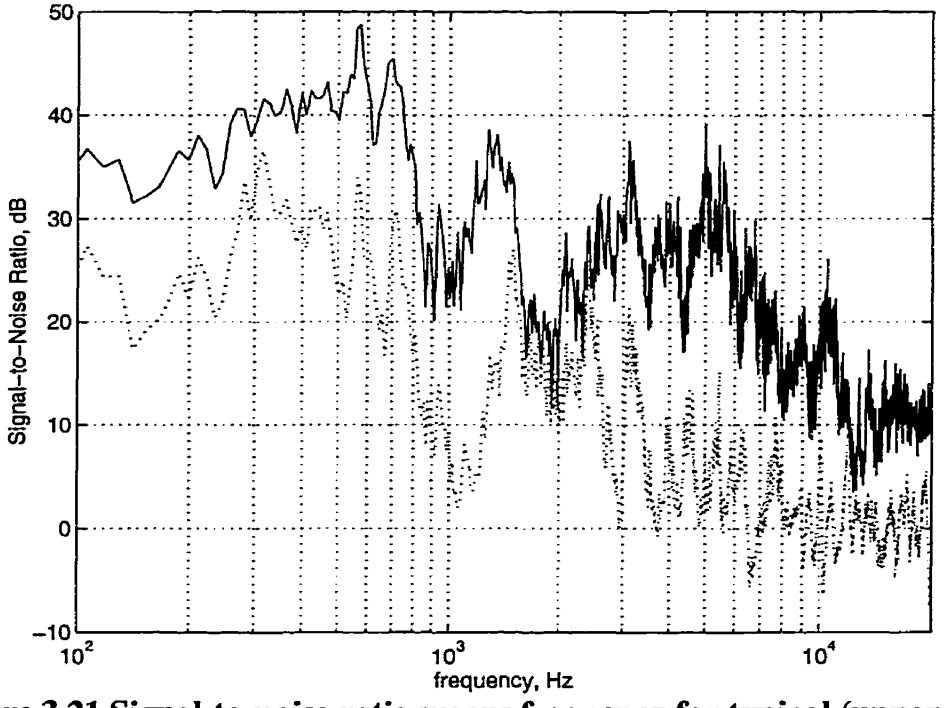


Figure 3.21 Signal-to-noise ratio versus frequency for typical (upper curve) and worst case (lower curve) camera noise.

The results shown in Figure 3.21 suggest that a single value description, such as signal-to-noise ratio, does not adequately describe the relation between the speech and the interfering camera noise due to the time-varying nature of speech. It is important to note that these curves represent the signal-to-noise ratios at only one measurement angle and only one configuration of the camera. Also, these measurements are for the NFB camera which is significantly quieter than the IMAX cameras.

3.6 Measurements of the IMAX Cameras

In this section, measurements of the noise produced by a broad range of IMAX cameras are described. The IMAX corporation produces a specialized type of motion picture using a large-screen format. IMAX films are projected onto very large screens (several sto-

ries high) in order to fully encompass the viewer's peripheral field of view. In order to retain resolution and picture quality, IMAX uses a wide gauge film (70 mm), and requires a sophisticated transport mechanism to move the film at very high speeds. As a result, an IMAX camera has only 1.5 to 3 minutes of film on a reel, and so reel changes occur frequently during filming. Furthermore, IMAX cameras are noisier than conventional cameras due to their complex mechanical workings. IMAX cameras are also larger, heavier, and more expensive than conventional cameras. For example, the IMAX-3D camera (of which there are only 2) costs US$ 2 million and requires 4 people to carry it. Therefore, a blimp which would add to the size and weight of the camera is a highly undesirable approach to reducing it's noise level.

IMAX films have consisted primarily of documentary style films based on topics such as; the N.A.S.A. space shuttle and Russian MIR space station missions, underwater footage of the Titanic, an expedition to the summit of Mount Everest, etc. The restriction in subject matter is due in part to the high cost of ADR associated with IMAX films. Due to the high noise levels produced by the cameras, a conventional movie (i.e., with actors and dialogue) filmed in the IMAX format requires that all dialogue be dubbed. Therefore, IMAX films would benefit greatly from a successful camera noise reduction system. Despite the somewhat restricted subject matter of IMAX films, they are viewed by more than 60 million people each year.

3.6.1 Recordings Provided by IMAX

The IMAX corporation provided the author with 2 sets of recordings. These included some of the recordings used by SAIC in their research collaboration with IMAX (see Chapter 2). The first recording consisted of a professional actor reciting dialogue while being filmed with a *standard* IMAX camera (model MSM 9801). This recording is useful in that it provides a recording of dialogue in a real-world setting with which to test the noise reduction algorithms.

Using this recording, measurements were made to examine the changes in the spectrum of this camera noise over time. The results are plotted in Figure 3.22 which shows the average power spectrum measured at 4 intervals in the recording. It can be seen that, like the NFB camera, the power spectrum of the IMAX MSM 9801 camera is relatively constant within a reel of film. Somewhat larger variations in the measured power spectra are seen for frequencies above 2 kHz.

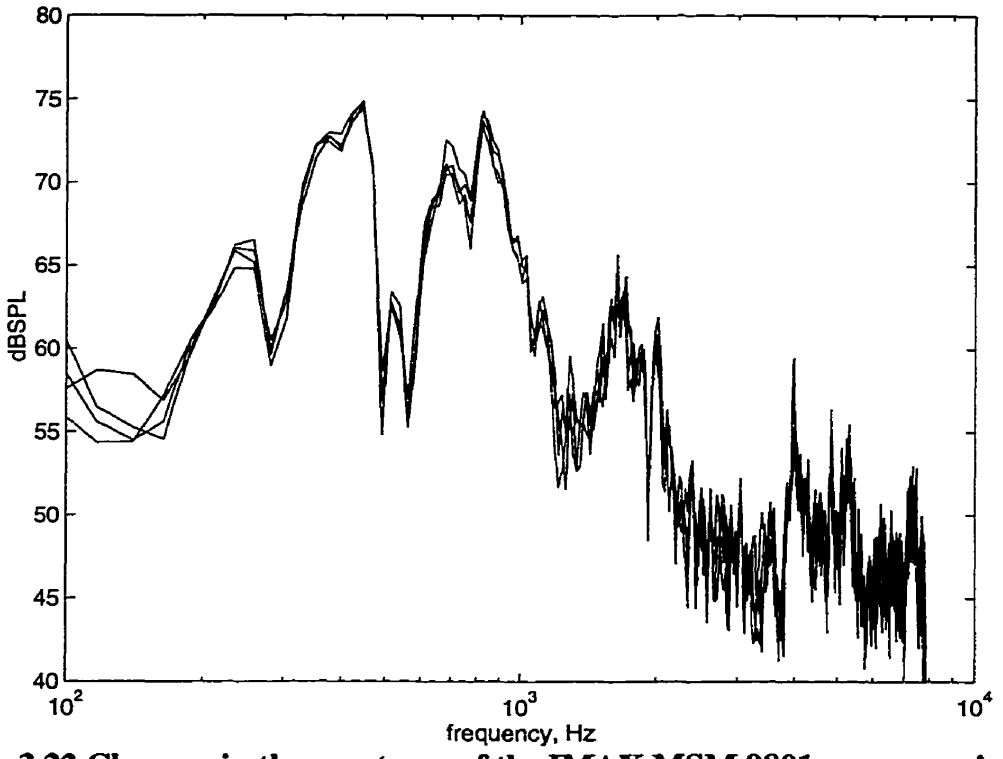


Figure 3.22 Changes in the spectrum of the IMAX MSM 9801 camera noise over time.

The second set of recordings consisted of numerous sequences recorded during the docking of the N.A.S.A. space shuttle with the Russian MIR space station. The recordings consist of 30 s sequences of the astronauts performing their duties during a space mission. These recordings contain high levels of background noise and exhibit strong multipath reflections. These recordings provide a good example of the need for a camera noise reduction system, since there is no possibility of ADR with the astronauts. The camera noise on these recordings is from the model MKII IMAX camera which is relatively small and lightweight. Due to the high level of the background noise and the fact that there is dialogue throughout the recordings, measurements of the power spectrum of the camera noise over time are not presented here.

3.6.2 IMAX-3D Camera

The author was given access to the IMAX-3D camera to make recordings and measurements. Recordings were made at two indoor locations and one outdoor location in order to provide a variety of real-world environments. All recordings were made with 2 professional quality microphones recording onto separate channels of a DAT recorder. One microphone was placed at the talker's location, while the other was placed next to the cam-

era. Recordings of the camera alone (i.e., no speech signal) and of camera with dialogue were made. This allowed some of the measurements conducted at the NFB to be repeated for the IMAX-3D camera.

Measurements were made to examine the changes in the spectrum of the IMAX-3D camera noise over time. The results are plotted in Figure 3.23 which shows the average power spectrum measured at 5 intervals in the recording. Again, the power spectrum is relatively constant within a reel of film thus further supporting the findings for the other cameras, as well as the model for camera noise proposed earlier.

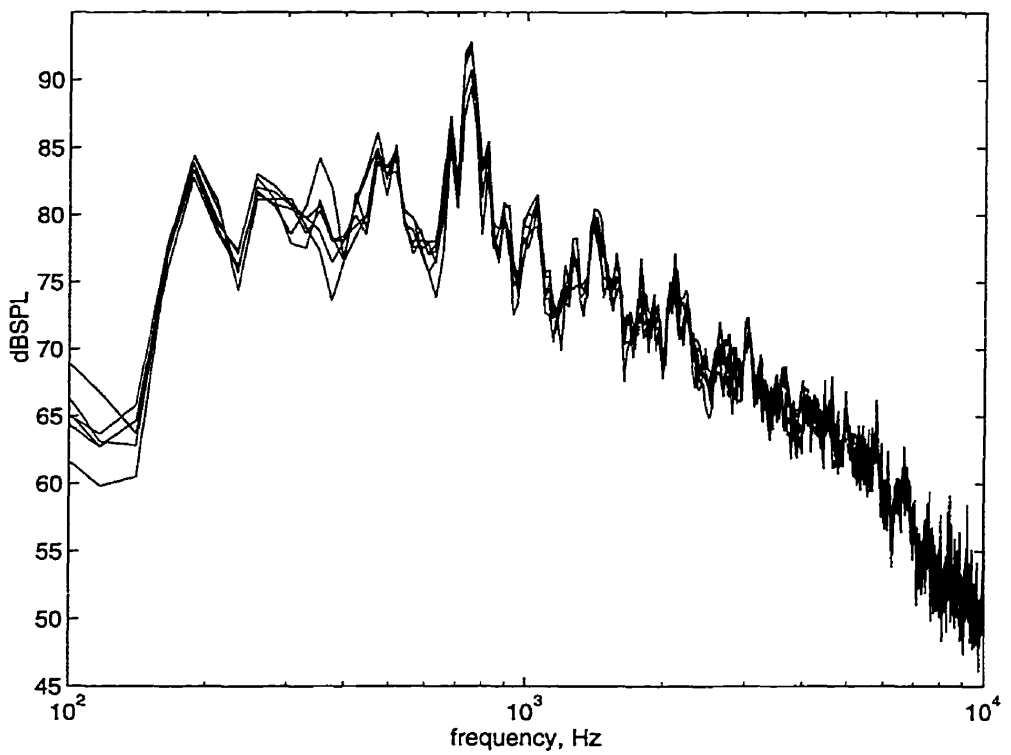


Figure 3.23 Changes in the spectrum of the IMAX-3D camera noise over time.

3.7 The Camera as a Distributed Noise Source

The varying directivity of the camera noise with frequency as seen in Section 3.5.2 suggests that the camera may be a distributed noise source. This observation is supported by the author's experience when making the recordings of both the NFB and the IMAX-3D cameras. By moving a microphone to different positions around the cameras, it was possible to hear different components of the noise being emphasized.

In this section, the distributed nature of the camera noise is examined. That is, we seek to demonstrate that the camera noise $n(k)$ is composed of P components distributed

in space,

$$n(k) = \sum_{i=1}^{P} n_i(k) \tag{3.1}$$

where $n_i(k)$ represent the various noise components such as; the opening and closing of the shutters, the motion of the film through the camera, the motor(s) driving the mechanism.

We begin by defining the magnitude-squared coherence $C_{yx}(\omega)$ between the observed stationary signals $x(k)$ and $y(k)$ at the two microphones,

$$C_{yx}(\omega) \equiv \frac{|S_{yx}(\omega)|^2}{S_{yy}(\omega)S_{xx}(\omega)} \; . \tag{3.2}$$

$S_{yx}(\omega)$ is the complex cross-power spectrum,

$$S_{yx}(\omega) = \sum_{l=-\infty}^{\infty} r_{yx}(l)e^{-j\omega l} \tag{3.3}$$

where

$$r_{yx}(l) = E[y(k)x(k-l)]$$

is the cross-correlation. $S_{xx}(\omega)$ and $S_{yy}(\omega)$ are the power spectra of $x(k)$ and $y(k)$ respectively as defined by,

$$S_{aa}(\omega) = \sum_{l=-\infty}^{\infty} r_{aa}(l)e^{-j\omega l} \tag{3.4}$$

where

$$r_{aa}(l) = E[a(k)a(k-l)] \; . \tag{3.5}$$

It can be shown that, given a point sound source (i.e., non-distributed) in a noise-free room, the magnitude-squared coherence $C_{yx}(\omega)$ between the signals measured at two locations in the room will be equal to 1. If however, the sound source is distributed in space, then $C_{yx}(\omega)$ will be equal to some value less than unity [46]. Therefore, $C_{yx}(\omega)$ can be used as a measure of the degree of distribution of a sound source.

The magnitude-squared coherence, $C_{yx}(\omega)$ measured for the NFB camera is shown as the solid curve in Figure 3.24. This curve was obtained by calculating $C_{yx}(\omega)$ of the camera noise recorded at 2 microphones. One microphone was at 0° and the other microphone was at 90°. It can be seen that, at most frequencies, the magnitude-squared coherence is well below 1.0. In fact, for most frequencies, $C_{yx}(\omega)$ is below 0.6. Thus, it can be

concluded that the camera is indeed a distributed noise source. The importance of this result will be seen in Chapter 4 where it will be shown that, due to the relatively low $C_{yx}(\omega)$, the performance of noise reduction schemes based on adaptive noise cancellation or blind signal separation is severely limited.

$C_{yx}(\omega)$ is sensitive to any misalignment of the signals $y(k)$ and $x(k)$. Therefore, in calculating the values for $C_{yx}(\omega)$ in Figure 3.24, $y(k)$ and $x(k)$ were shifted in time relative to each other until the maximum value for $C_{yx}(\omega)$ was obtained, thus eliminating any misalignment effects. This method was used for all measurements of $C_{yx}(\omega)$ reported in this thesis.

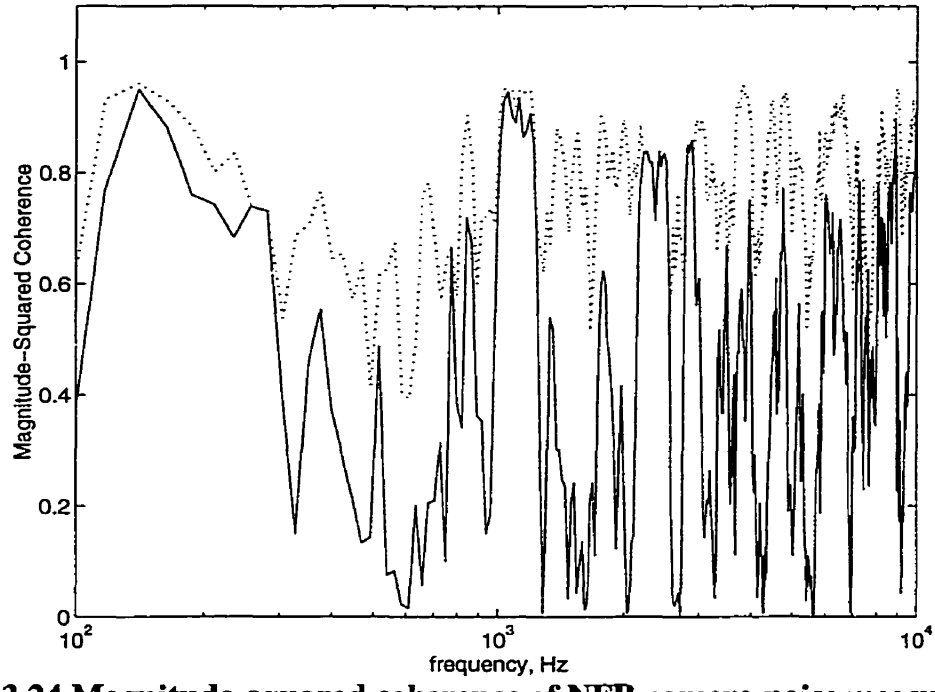


Figure 3.24 Magnitude-squared coherence of NFB camera noise measured at 2 microphones.

One might ask whether $C_{yx}(\omega)$ would increase if the microphones had been placed at different locations around the camera. To address this question, measurements were taken at 30 different locations around the camera. All of the measurements had $C_{yx}(\omega)$ values similar to that plotted in Figure 3.24. The dotted line in Figure 3.24 represents the maximum value of $C_{yx}(\omega)$ found at each frequency across all of the measurements. The results show that the low values of $C_{yx}(\omega)$ are not due to the location of the measurement microphones, and so the conclusion that the camera is a distributed noise source is further strengthened.

While recording the IMAX-3D camera, it seemed evident that, due to its large size and complex workings, it was acting as a distributed noise source. This observation was tested by measuring the magnitude-squared coherence $C_{yx}(\omega)$ between the two channels of the recording of the camera noise alone (i.e., without a speech signal). The result is plotted in Figure 3.25. It can be seen that $C_{yx}(\omega)$ is quite low for most frequencies, and exceeds 0.5 for only a few individual frequencies. The result of this measurement confirms the observation that the IMAX-3D camera is a distributed noise source. Therefore, noise reduction techniques based on adaptive noise cancellation or blind signal separation are not likely to perform adequately for this camera. This conclusion is confirmed in Chapter 4.

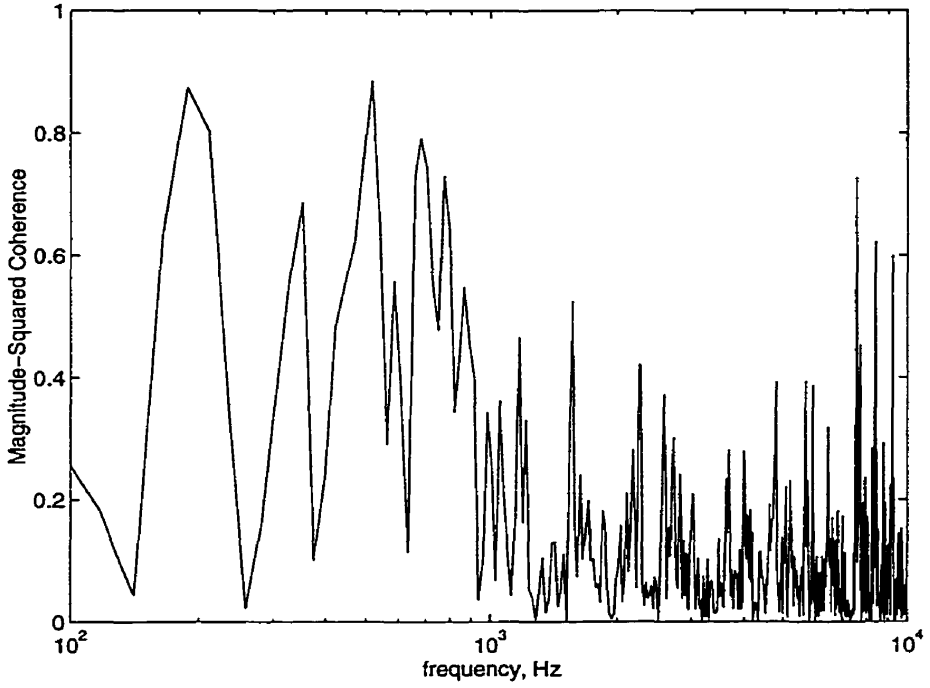


Figure 3.25 Magnitude-squared coherence of IMAX-3D camera noise measured at 2 microphones.

It is interesting to compare the results in Figure 3.25 to the results for the NFB camera shown in Figure 3.24. It can be seen that overall, $C_{yx}(\omega)$ for the IMAX-3D camera tends to be significantly lower than for the NFB camera. This suggests that the IMAX-3D camera is the more distributed noise source. This conclusion is sensible given that the IMAX-3D camera is much larger than the NFB camera, and it has more complex inner workings. Moreover, the IMAX-3D camera uses two reels of film (one for each lens) and so it effectively consists of two cameras in one.

3.8 A Model of Camera Noise

In considering various schemes for reducing camera noise, it is helpful to have a model of the composition of the noise source. In this section, a simple model is developed which decomposes camera noise $n(k)$ into two fundamental components: a periodic component and a cyclical random component. That is, the model is of the form

$$n(k) = p(k) + c(k), \tag{3.6}$$

where $p(k)$ represents the periodic component and $c(k)$ represents the cyclical random component.

The time domain waveforms (see Figures 3.4 and 3.5) as well as the spectrogram of Figure 3.8 clearly indicate that camera noise is not a stationary random process. Rather, camera noise consists of regularly spaced pulses which are similar to each other. However, differences between the pulses are also clearly visible. There is a random component to the camera noise which is most readily seen between the pulses. Figure 3.19, Figure 3.20, and Figure 3.22 show that the power spectrum of the camera noise (within a reel of film) remains relatively unchanged over time. That is, from frame to frame, the power spectrum of the camera noise does not change dramatically. However, the spectrogram of Figure 3.8 clearly demonstrates that the camera noise varies significantly within a frame. Therefore, the model should reflect these two aspects of camera noise.

In deriving a model of camera noise we begin by considering the periodic component $p(k)$. Let $Q(k)$ be the sum of all the camera noise components $q_i(k)$ which repeat from frame to frame

$$Q(k) = \begin{cases} \sum_i q_i(k) & ; k=1,2,\ldots,T \\ 0 & ; \textit{elsewhere} \end{cases}, \tag{3.7}$$

where T is the period. The periodic component $p(k)$ of the camera noise $n(k)$ is then defined as,

$$p(k) = \sum_{l=-\infty}^{\infty} Q(k+lT). \tag{3.8}$$

The periodic component of the camera noise is likely due to the opening and closing of the shutters, the movement of the sprockets which guide the film through the camera, as well as the motor(s) driving the overall mechanism.

The second component of the camera noise is the cyclical random component, $c(k)$ which is modeled as a zero-mean stationary random process whose amplitude is modu-

lated by a *gating function* $G(k)$. This component is described as a cyclical random process since it consists of a random process whose statistics fluctuate within each frame, but repeat from frame to frame. That is, the statistics of the noise are cyclical. The gating function is defined as,

$$G(k)=\begin{cases}\alpha(\omega) & ; k=1,2,...,\tau(\omega) \\ \beta(\omega) & ; k=\tau(\omega)+1,...,T\end{cases} . \tag{3.9}$$

The parameter $\alpha(\omega)$ determines the level of $c(k)$ during the peak (pulse) portion of the camera noise, while $\beta(\omega)$ controls the level between the pulses. $\tau(\omega)$ is chosen to be equal to the duration of the pulse portion of the camera noise. The three parameters, $\alpha(\omega)$, $\beta(\omega)$, and $\tau(\omega)$ are all assumed to be a function of frequency to reflect the results of Figure 3.8. Given this, the cyclical random component $c(k)$ of the camera noise is defined as,

$$c(k)=v(k)\cdot\sum_{l=-\infty}^{\infty}G(k+lT) , \tag{3.10}$$

where $v(k)$ is a zero-mean stationary random process.

This component of the camera noise is composed of several sources including; the random movement of the film as it passes through the camera, the vibrations of the body of the camera, and all noise sources which are not directly related to the film rate of the camera.

The camera noise $n(k)$ can now be defined in terms of the two components $p(k)$ and $c(k)$,

$$n(k)=\sum_{l=-\infty}^{\infty}[Q(k+lT)+v(k)G(k+lT)] . \tag{3.11}$$

Equation (3.11) provides a simple mathematical model of the camera noise. The limitations of this simple model will be seen in later chapters. In Chapter 4, methods which seek to exploit the periodic component $p(k)$ are examined, while Chapter 6 addresses the cyclical random component, $c(k)$.

3.9 Summary

In this chapter a series of acoustic measurements designed to characterize camera noise were described. Measurements at the NFB showed that a significant component of the noise is related to the film rate of the camera. However, not all of the noise is directly related to the film rate. Based on these findings, a simple mathematical model of the

camera noise was derived which includes a periodic component and a cyclical random component.

It was found that the camera is a distributed noise source, and so the performance of noise reduction schemes based on adaptive noise cancellation or blind signal separation will be compromised.

The type of lens used on the camera was found to have a small effect on the noise, whereas the film stock had a much larger effect. Somewhat surprisingly, the location within a reel of film had virtually no effect on the camera noise.

Measurements were also made from recordings of 3 models of IMAX cameras. While the level of the noise from the IMAX cameras is significantly higher than for the NFB camera, the main findings of the NFB measurements are supported. That is, the cameras act as distributed noise sources and the power spectrum of the noise is relatively constant for a given reel of film.

The measurements described in this chapter demonstrate that camera noise can be quite variable due to several factors. Therefore, a static processing approach can not adequately address the problem of camera noise, and a successful noise reduction scheme must be adaptive. Also, the main characteristics of camera noise appear to be common to the 4 cameras which were examined. These represent a broad range of camera types.

4. NOISE REDUCTION USING ADAPTIVE FILTERING METHODS

4.1 Introduction

In this chapter adaptive filtering methods are examined as a potential means of reducing camera noise. Specifically, the adaptive noise cancellation (ANC) method and the more recently developed blind signal separation techniques are explored. In their basic forms both of these adaptive filtering techniques require at least two input signals and thus violate the single-input requirement for a practical scheme for reducing camera noise. Nonetheless, these techniques are worth exploring since they can provide valuable insight into the noise reduction problem.

The chapter begins with a brief overview of Wiener filter theory and the concept of optimal filtering of stationary random processes. The ANC problem is described with an emphasis on the Widrow-Hoff LMS algorithm. Results of simulations to examine the use of the LMS-based ANC method for reducing camera noise are presented. The underlying principles of blind signal separation techniques are provided and methods based on second order statistics are described in some detail.

The similarities between the noise cancellation problem and the blind signal separation problem are highlighted. It will be seen that the signal separation problem can be viewed as the more general case of noise cancellation. Furthermore, the ANC method is shown to be a special case of some signal separation algorithms. Included in this chapter is a discussion of some of the factors which can limit the performance of these adaptive filtering methods for this application.

It was originally believed at the start of this research work that an LMS-based ANC method was the obvious means of reducing camera noise. In an effort to overcome the two (or multi) input requirement of the adaptive filtering approach, a variation to the ANC method is proposed wherein the reference input signal is synthesized, thus effectively creating a single-input ANC system. However, in order to provide a useful degree of noise cancellation, the synthesized reference input approach requires a high correlation between the individual pulses of the camera noise. Unfortunately, it will be seen that the inter-pulse correlation is relatively low

4.2 *Wiener Filters*

In this section the discrete-time Wiener filter is examined for the case of a real-valued time series. We will restrict the discussion to the formulation of a finite-duration impulse response (FIR) filter. Wiener theory provides the means to determine the coefficients of the filter which minimizes the mean-squared error between the filter output and some desired signal.

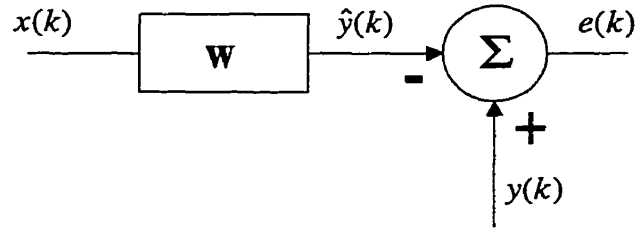


Figure 4.1 Block diagram of Wiener filter.

Consider the linear transversal filter **w** of order $N-1$ shown in Figure 4.1. The input $x(k)$ to the filter is assumed to be wide-sense stationary with zero mean. The tap weights for the filter are w_i, i=1,2, ...,N. The output of the filter, $\hat{y}(k)$ represents an estimate of some desired signal $y(k)$, and is obtained through the convolutional sum

$$\hat{y}(k)=\sum_{l=1}^{N} w_l x(k-l+1). \tag{4.1}$$

The goal is to determine the coefficients w_i, i=1,2, ...,N such that the difference between the desired signal $y(k)$, and the estimate of the desired signal $\hat{y}(k)$, is minimized. That is, we want to somehow minimize the estimation error, $e(k)$ which is defined as,

$$e(k)=y(k)-\hat{y}(k). \tag{4.2}$$

Wiener theory uses the *minimum mean-square error* as the criteria for optimizing the filter. Specifically, the filter coefficients are chosen so as to minimize the cost function $J(w)$ defined as

$$J(w)=E[e^2(k)]. \tag{4.3a}$$

$$=E[(y(k)-\hat{y}(k))^2] \tag{4.3b}$$

Thus, $J(w)$ represents the mean-squared error.

We now consider the cost function in terms of the desired signal, the input signal, and the filter tap weights. Define the tap weight vector

$$\mathbf{w}^T=[w_1(k),w_2(k),...,w_N(k)] \tag{4.4}$$

and the input vector

$$\mathbf{x}^T(k)=[x(k),x(k-1),...,x(k-N+1)]. \tag{4.5}$$

The convolutional sum of (4.1) can now be expressed as the inner product of the tap weight vector and the input vector

$$\hat{y}(k)=\mathbf{w}^T\mathbf{x}. \tag{4.6}$$

The mean-squared error $J(w)$ now becomes

$$\begin{aligned} J(w)&=E[(y(k)-\mathbf{w}^T\mathbf{x})^2] \\ &=E[y(k)^2]-2\mathbf{w}^TE[\mathbf{x}(k)y(k)]+\mathbf{w}^TE[\mathbf{x}(k)\mathbf{x}^T(k)]\mathbf{w}. \end{aligned} \tag{4.7}$$

Note that the error is a second order function and thus has a unique minimum. Our goal is to find the value of **w** which gives this minimum value.

Denoting the cross-correlation vector between the input and the desired signal as

$$\mathbf{r}_{xy}=E[\mathbf{x}(k)y(k)] \tag{4.8}$$

and the correlation matrix of the input signal as

$$\mathbf{R}_{xx}=\begin{bmatrix} r_{xx}(0) & r_{xx}(1) & r_{xx}(2) & \cdots & r_{xx}(N-1) \\ r_{xx}(1) & r_{xx}(0) & r_{xx}(1) & \ddots & r_{xx}(N-2) \\ r_{xx}(2) & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ r_{xx}(N-1) & r_{xx}(N-2) & \cdots & \cdots & r_{xx}(0) \end{bmatrix}, \tag{4.9}$$

where

$$r_{xx}(l)=E[x(k)x(k-l)]\ . \tag{4.10}$$

The mean-squared error $J(w)$ can now be written in terms of these new expressions to give

$$J(w)=r_{yy}(0)-2\mathbf{w}^T\mathbf{r}_{xy}+\mathbf{w}^T\mathbf{R}_{xx}\mathbf{w}. \tag{4.11}$$

The gradient of the mean-squared error $J(w)$ with respect to the tap weight vector is given by

$$\nabla=\frac{\partial}{\partial\mathbf{w}}J(w). \tag{4.12a}$$

$$=-2\mathbf{r}_{xy}+2\mathbf{w}^T\mathbf{R}_{xx}. \tag{4.12b}$$

Setting this to zero, we obtain the discrete form of the Wiener-Hopf equation, or the so-called *normal* equation

$$\mathbf{w}_{opt} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xy}. \tag{4.13}$$

Therefore, $\mathbf{w}_{opt}$ represents the coefficients of the minimum mean-squared error Wiener filter. Substituting this optimal value for $\mathbf{w}$ into (4.11) we get

$$\begin{aligned} J(w) &= r_{yy}(0) - \mathbf{w}^T\mathbf{r}_{xy} \\ &= r_{yy}(0) - r_{\hat{y}\hat{y}}(0) \end{aligned} \tag{4.14}$$

which is the minimum mean-squared error obtained with the Wiener filter.

4.3 Adaptive Noise Cancellation

In this section, we describe the basic aspects of adaptive noise cancellation as applied to the problem of reducing camera noise. It should be noted that the acoustic nature of the signals introduces certain practical limitations which will be discussed later. We begin by considering the non-adaptive structure shown in Figure 4.2.

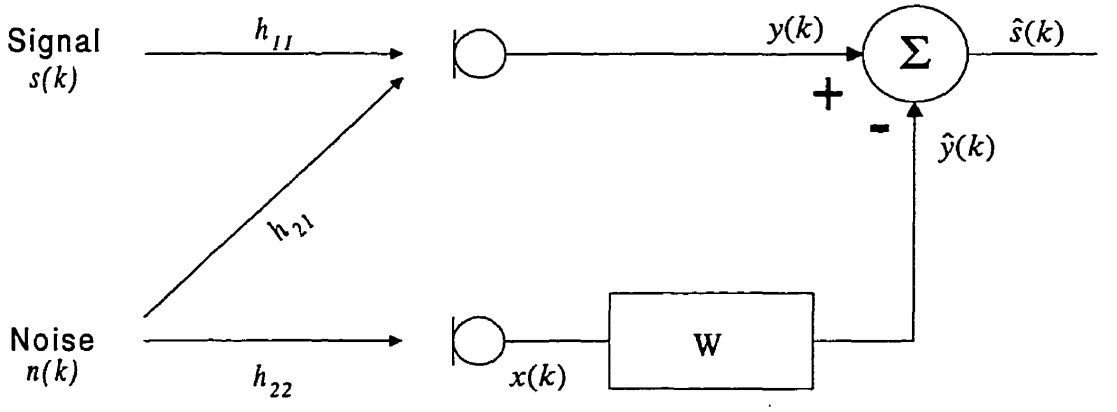


Figure 4.2 Block diagram of non-adaptive noise cancellation system.

The figure shows a desired signal $s(k)$ which is corrupted by an additive noise $n(k)$ and picked up at a receiver to form the *primary* input signal $y(k)$. Both $s(k)$ and $n(k)$ are assumed to be wide-sense stationary. As a result of the paths through which they travel in the room, the signal $s(k)$ is convolved with the acoustic impulse response h_{11} and $n(k)$ is convolved with h_{21}. That is,

$$y(k) = s(k) * h_{11} + n(k) * h_{21} \tag{4.15a}$$

$$= \sum_{l=0}^{\infty} h_{11}(l)s(l-k) + \sum_{m=0}^{\infty} h_{21}(m)n(m-k) \tag{4.15b}$$

In the adaptive noise cancellation method, it is assumed that a second *reference* input $x(k)$ is available which contains a signal which is in some way correlated with $n(k)*h_{21}(k)$. From Figure 4.2, we see that for the case considered here,

$$x(k)=n(k)*h_{22} \tag{4.16a}$$

$$=\sum_{l=0}^{\infty}h_{22}(l)n(l-k) \tag{4.16b}$$

The noise cancellation method now consists of filtering the reference signal $x(k)$ through $\mathbf{w}$ to yield $\hat{y}(k)$ such that the subtraction of $\hat{y}(k)$ from $y(k)$ gives the optimal estimate of the desired signal $s(k)$. In mathematical terms, we want to find $\mathbf{w}$ such that

$$\min\{J(w)\}=\min\{E[f(y(k)-\hat{y}(k))]\} \tag{4.17}$$

where $f(\ )$ is some function which is used as the criterion for minimizing $J(w)$. If we choose to minimize $J(w)$ in the mean-squared sense, then we obtain

$$\min\{J(w)\}=\min\{E[(y(k)-\hat{y}(k))^2]\}. \tag{4.18}$$

We note that this is the same cost function that was used in (4.3) to derive the optimal Wiener filter. Therefore, for our noise cancellation problem, the optimal estimate of the desired signal $s(k)$ is obtained by setting $\mathbf{w}$ equal to the optimal value, $\mathbf{w}_{opt}$ found in the normal equation (4.13). Equivalently, we can say that $\mathbf{w}_{opt}$ is the value of $\mathbf{w}$ which minimizes (4.18).

In general, we cannot assume that the filter $\mathbf{w}$ will be operating in a stationary environment and so, $\mathbf{w}_{opt}$ may be continuously changing. To account for the changing $\mathbf{w}_{opt}$, an adaptive filter is used. The problem then becomes one of designing the adaptive filter so that it tracks the changes in the operating environment and remains as close as possible to $\mathbf{w}_{opt}$. The basic Adaptive Noise Cancellation scheme is illustrated in Figure 4.3.

The reference input $x(k)$ is filtered by the adaptive filter $\mathbf{w}(k)$ to yield the sequence $\hat{y}(k)$ which is an estimate of the primary signal $y(k)$. $\hat{y}(k)$ is then subtracted from $y(k)$ to give $e(k)$ which is the error signal used by the adaptation algorithm to find the optimal filter $\mathbf{w}(k)$ at time k.

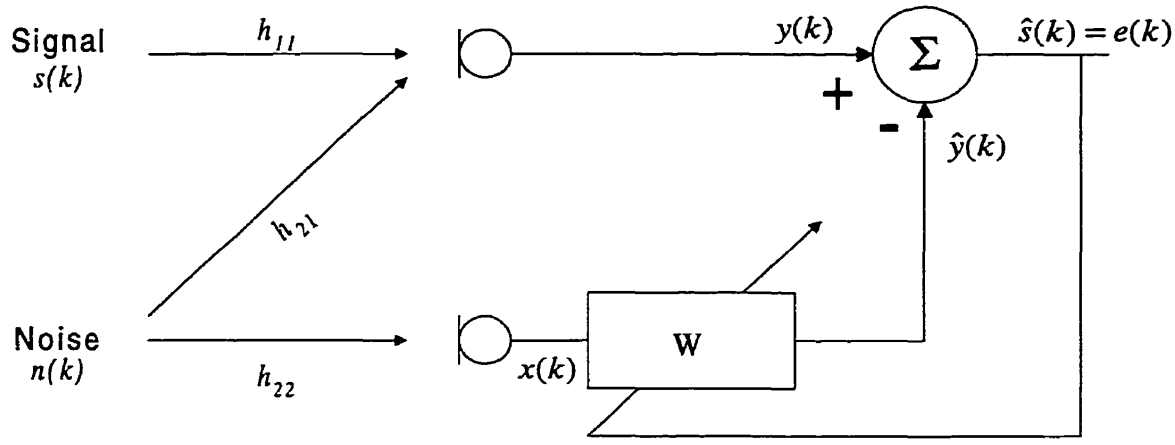


Figure 4.3 Block diagram of adaptive noise cancellation system.

Widrow *et al.* [7] showed that minimizing $e(k)$ is equivalent to finding the best estimate of $s(k)$. This is intuitively satisfying since $x(k)$ is assumed to be correlated with only the undesired (i.e., noisy) portion of $y(k)$. Therefore, by removing as much as possible of the portion of $y(k)$ associated with $x(k)$, (i.e., by minimizing $e(k)$) we are left with an optimal estimate of $s(k)$. As such, the error signal $e(k)$ is in fact, the estimate of the desired signal.

Solving the normal equation (4.13) directly in order to find $\mathbf{w}_{opt}$ is a computationally demanding process (especially for higher order filters) since it includes inverting the correlation matrix $\mathbf{R}_{xx}$. We therefore seek a simplified means of obtaining an estimate of $\mathbf{w}_{opt}$. There are many adaptation algorithms which can be used to update the coefficients of the adaptive filter. The choice of adaptation algorithm will have a profound effect on the performance of the adaptive noise cancellation system. In practice, the choice of algorithm is determined by how quickly and accurately it tracks changes in the operating environment, its computational complexity, its robustness against instability, and the types of input signals expected.

4.3.1 The Widrow-Hoff LMS Algorithm

The most commonly used adaptation scheme is the Widrow-Hoff least-mean-square (LMS) algorithm which is based on the method of steepest descent [8,47,7]. One of the key features of the LMS algorithm is its low computational complexity. It does not require explicit calculation of the correlation matrix $\mathbf{R}_{xx}$ or the cross-correlation vector $\mathbf{r}_{xy}$. Furthermore, it does not require matrix inversion. The LMS algorithm is derived below.

We begin by examining the method of steepest descent. Assume that $\mathbf{w}(k)$ is a linear transversal filter of order N-1. To find the optimal value of $\mathbf{w}(k)$ using the method of

steepest descent, we use the recursive relation

$$\mathbf{w}(k+1)=\mathbf{w}(k)+\frac{1}{2}\mu[-\nabla(k)] \tag{4.19}$$

$$=\mathbf{w}(k)-\frac{1}{2}\mu\frac{\partial J(w)}{\partial w} \tag{4.20}$$

where

$$\mathbf{w}^T(k)=[w_1(k),w_2(k),...,w_N(k)]. \tag{4.21}$$

The estimate of $\mathbf{w}_{opt}$ at time k+1 is determined by the estimate at time k and the gradient. Substituting the solution for ∇ derived in (4.12) into (4.20) we get

$$\mathbf{w}(k+1)=\mathbf{w}(k)+\mu[\mathbf{r}_{xy}-\mathbf{R}_{xx}\mathbf{w}(k)]. \tag{4.22}$$

In the LMS algorithm, instantaneous estimates of $\mathbf{r}_{xy}$ and $\mathbf{R}_{xx}$ are used. We define the instantaneous estimates as

$$\tilde{\mathbf{r}}_{xy}(k)=\mathbf{x}(k)y(k) \qquad \text{and} \tag{4.23}$$

$$\tilde{\mathbf{R}}_{xx}(k)=\mathbf{x}(k)\mathbf{x}^T(k)\ . \tag{4.24}$$

Replacing the correlation matrix $\mathbf{R}_{xx}$ and the cross-correlation vector $\mathbf{r}_{xy}$ in (4.22) by their instantaneous estimates, we get

$$\mathbf{w}(k+1)=\mathbf{w}(k)+\mu[\mathbf{x}(k)y(k)-\mathbf{x}(k)\mathbf{x}^T(k)\mathbf{w}(k)] \tag{4.25a}$$

$$=\mathbf{w}(k)+\mu\mathbf{x}(k)[y(k)-\mathbf{x}^T(k)\mathbf{w}(k)]. \tag{4.25b}$$

The result of (4.25) can be written as

$$e(k)=y(k)-\mathbf{x}^T(k)\mathbf{w}(k) \tag{4.26}$$

$$\mathbf{w}(k+1)=\mathbf{w}(k)+\mu\mathbf{x}(k)e(k). \tag{4.27}$$

Equations (4.26) and (4.27) form the basis of the LMS algorithm. The variable μ is the adaptation step size which determines the rate at which $\mathbf{w}(k)$ converges towards the optimal solution. Comprehensive discussions regarding the ANC method and the LMS algorithm can be found in [8,9,10,48].

4.3.2 Limitations of the LMS Algorithm

While the LMS algorithm provides a computationally efficient means of steering the adaptive filter towards its optimal solution, it has several limitations due to its simplicity. In the ANC application the task of the LMS algorithm is to adapt the transversal filter $\mathbf{w}(k)$ so as to obtain the maximum amount of noise reduction. If we assume (as is often done) that h_{11}=1 and h_{22}=1, then it can be seen from Figure 4.3 that optimal performance

occurs when the taps of the filter are equal to the impulse response h_{21} (i.e., $\mathbf{w}_{opt} = h_{21}$). That is, the adaptive filter is attempting to predict the acoustic path traveled by the noise $n(k)$ to the primary microphone.

One of the main limitations of the LMS algorithm is its potentially slow rate of convergence. The rate of convergence determines the time required for $\mathbf{w}(k)$ to reach a sufficiently close approximation to h_{21} and provide a sufficient degree of noise reduction. The rate of convergence also affects the ability of the adaptive filter to track changes in h_{21}.

A key factor in determining the rate of convergence of the LMS algorithm is the eigenvalue spread of the correlation matrix $\mathbf{R}_{xx}$ [8,10,49]. The eigenvalue spread $\chi(\mathbf{R})$ is defined as

$$\chi(\mathbf{R}) = \frac{\lambda_{max}}{\lambda_{min}} \tag{4.28}$$

where λ_{max} and λ_{min} are the largest and smallest eigenvalues of $\mathbf{R}_{xx}$. In matrix terms, the eigenvalue spread is related to the condition number of the correlation matrix $\mathbf{R}_{xx}$ [50]. When the eigenvalue spread is large, the LMS algorithm requires a greater number of iterations (i.e., more time) in order to converge.

The effect of the eigenvalue spread can be seen directly in the adaptation stepsize parameter μ of (4.27). A larger value of μ allows the LMS algorithm to converge more quickly than a smaller value of μ. However, if μ is too large, the algorithm will not converge. The largest value of μ which will still provide a convergent system, is determined by the maximum eigenvalue λ_{max} of the correlation matrix $\mathbf{R}_{xx}$. Specifically, a necessary condition for convergence in the mean is that μ must lie within the interval [8,9,10,48,49]

$$0 < \mu < \frac{2}{\lambda_{max}} . \tag{4.29}$$

The relation between λ_{max} and the rate of convergence is evident from (4.29). For a given value of μ, the rate of convergence is determined by the mode corresponding to the smallest eigenvalue, λ_{min}. A smaller λ_{min} implies a larger decay time of the mode, and hence a slower rate of convergence.

While a large value for μ implies a faster rate of convergence, it also implies a larger misadjustment M [8,9,10,49]. That is, a larger value of μ limits the accuracy with which $\mathbf{w}(k)$ converges to $\mathbf{w}_{opt}$. The relation between the misadjustment M and the adaptive stepsize μ can be approximated by

$$M \approx J_{\min}(1+\frac{\mu N r_{xx}(0)}{2}) \tag{4.30}$$

for a transversal filter of order N–1. Therefore, there is an inevitable tradeoff between the speed with which the adaptive filter converges and the amount of noise reduction that an LMS-based ANC system can achieve. Moreover, the performance of the LMS-based ANC system will be directly influenced by the statistics of the noise to be canceled.

Other adaptive algorithms exist which do not suffer from the limitations of the LMS algorithm but are more (often dramatically) computationally demanding. One class of adaptive filters which can offer and improved rate of convergence while maintaining a reasonable level of computational complexity are the transform domain LMS algorithms. The concept of adaptive filtering in the frequency domain was proposed by Dentino *et al.* [51]. The theory behind transform domain adaptive filtering was further developed by Narayan *et al.* [52] and Lee and Un [53] including an improved understanding of the performance characteristics of this class of filters. Marshall *et al.* [54] studied the performance of the transform domain adaptive filter for a variety of orthogonal transforms. They found that considerably improved performance (i.e., rate of convergence) over LMS could be obtained for a broad class of input signals using the transform domain schemes.

The ratio of the maximum to the minimum eigenvalues of the correlation matrix $\mathbf{R}_{xx}$ is bounded by the ratio of the maximum to the minimum magnitudes of the power spectrum of $x(k)$ [55]. That is,

$$1 \le \frac{\lambda_{\max}}{\lambda_{\min}} \le \frac{\max|X(e^{j\omega})|^2}{\min|X(e^{j\omega})|^2}. \tag{4.31}$$

Given this, an approach to accelerate the rate of convergence of the adaptive algorithm is to somehow transform the input signal $x(k)$ into another signal $z(k)$ whose corresponding correlation matrix $\mathbf{R}_{zz}$ has a smaller eigenvalue spread. This can be achieved by performing the adaptive filtering in some orthogonal transform domain.

A block diagram of the transform domain adaptive filter is given in Figure 4.4 where it can be seen that the input signal vector $\mathbf{x}(k)$ is transformed into another vector $\mathbf{z}(k)$

$$\mathbf{z}^T(k) = [z_1(k), z_2(k), \ldots, z_N(k)] \tag{4.32}$$

using an orthogonal transformation

$$\mathbf{z}(k) = \mathbf{Q}\mathbf{x}(k). \tag{4.33}$$

The transform matrix $\mathbf{Q}$ is a unitary matrix of rank N and thus,

$$\mathbf{Q}\mathbf{Q}^T = \mathbf{I} \ . \tag{4.34}$$

The transformed input vector $\mathbf{z}(k)$ is multiplied by the transform domain tap weight vector $\mathbf{v}(k)$

$$\mathbf{v}^T(k) = [v_1(k), v_2(k), \ldots, v_N(k)] \tag{4.35}$$

to form the adaptive output $\hat{y}(k)$.

$$\hat{y}(k) = \mathbf{z}^T(k)\mathbf{v}(k) \tag{4.36}$$

The resulting error signal is

$$e(k) = y(k) - \hat{y}(k) \tag{4.37}$$

and the tap weight update equation is

$$v_i(k+1) = v_i(k) + 2\xi_i e(k) z_i(k) \quad i = 1,2,\ldots,N \tag{4.38}$$

where

$$\xi_i = \frac{\mu}{E[z_i^2(k)]} \tag{4.39}$$

is the adaptation stepsize parameter for the ith transform component and μ is a positive constant that controls the rate of convergence.

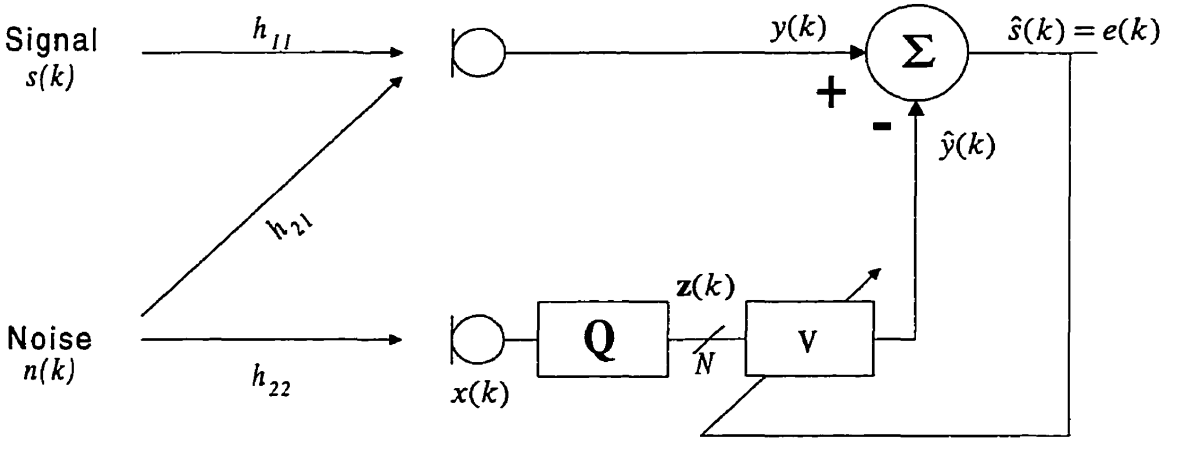


Figure 4.4 Block diagram of a transform domain LMS based ANC system.

The purpose of the transform $\mathbf{Q}$ is to decorrelate the input signal $x(k)$ thus minimizing the eigenvalue spread, and so the choice of an appropriate orthogonal transform is critical. It is well known that the Karhunen-Loéve transform (KLT) provides the optimal decorrelation of an input signal. The KLT is composed of the orthonormal eigenvectors of the input correlation matrix and is thus signal dependent. Because it is a signal dependent transform, the KLT is generally not practical for most applications.

Several researchers have studied the performance of transform domain LMS for a variety of time-invariant transforms. Lee and Un [53] evaluated the performance of transform domain LMS for a variety of orthogonal transforms assuming real-valued input data. The transforms that they investigated included; the discrete Fourier transform (DFT), the discrete cosine transform (DCT), the symmetric cosine transform (SCT), a fast KLT, and the discrete sine transform (DST). Marshall *et al.* [54] also studied the performance of several transforms including the DFT, the DCT, the Walsh-Hadamard transform (WHT), the discrete Hartley transform (DHT), and the Power-Of-Two transform (PO2) which was designed specifically for the transform domain LMS. They found that, in general, the DCT-LMS gave the best overall performance for speech signals. Narayan *et al.* [52] examined the performance of transform domain LMS using the DFT and the DCT. They concluded that, for speech applications, use of the DCT-LMS will provide faster convergence than the DFT.

The performance of the different algorithms depends on the orthogonalizing capabilities of the data-independent transform used to pre-whiten the input data. No general proof exists demonstrating the superiority of one transform over the others. However, Beaufays [56] recently proved that, for first-order Markov input signals, the eigenvalue spread after transformation by a DCT will always be less than for a DFT.

In recent years, transform domain LMS using a wavelet transform has been proposed [57,58,59,61]. A wavelet-based approach offers potential advantages over Fourier-based methods for cases when the time varying modes of the input signal are not adequately represented by a weighted sum of sinusoids. Hosur and Tewfik [59,60] showed that the wavelet transform LMS could provide better convergence performance than DCT-LMS. Attallah and Najim [61] found that a wavelet decomposition based on a regular subband tree yielded better results than one based on a dyadic tree. A comprehensive review of transform domain adaptive filtering can be found in the review paper by Shynk [62].

In this section, we focus on the LMS algorithm and its transform domain variants. Our interest in the LMS based algorithms is prompted by their low computational complexity. Of course, other non-LMS based adaptive algorithms exist. For example, recursive least-square (RLS) algorithms are known to exhibit near optimal convergence behavior, but suffer from high complexity and instability issues [49,8]. Fast least squares algorithms including those based on lattice structures do reduce the computational demand and provide stable performance. However due to the high computational demands of the camera noise problem, the use of these algorithms remains impractical.

As stated earlier, the task of the adaptive filter in the ANC system depicted in Figure 4.3 is to estimate the path h_{21}. In the camera noise application, h_{21} is the acoustic path from the camera to the microphone which is recording the actor's dialogue. Because it may be necessary in many instances for the ANC system to provide more than 20 or 30 dB of noise reduction, it may be necessary for $\mathbf{w}(k)$ to be a very high order filter (i.e., many taps). For example, in a room having a reverberation time of about 1.5 s, it may be necessary to cancel the multipath reflections of the first 500 ms of the acoustic impulse response h_{21} in order to achieve 20 dB of noise reduction. At the sampling rate of f_s=48000 Hz, an adaptive filter length of 24000 taps would be required. Even if it were deemed acceptable to reduce the sampling rate to 24000 Hz, the order of the adaptive filter would still exceed 10000 taps. Due to the large number of taps required, any practical implementation would likely necessitate using an LMS-based adaptation algorithm. Also, Boll and Pulsipher [63] found that an audible echo may be created in the output of an ANC system when using long filter lengths. The echo becomes more prominent as μ is increased since the accuracy with which $\mathbf{w}(k)$ converges is correspondingly reduced. That is, the individual taps of the filter wander about their optimal values, thus resulting in a large excess mean squared error.

4.3.3 Limitations of the ANC system

The ANC system depicted in Figure 4.3 represents and ideal case. In real-world applications, regardless of which algorithm is employed to adapt the filter, there are certain limitations of the ANC system which can constrain its performance. A somewhat more realistic scenario is shown in Figure 4.5.

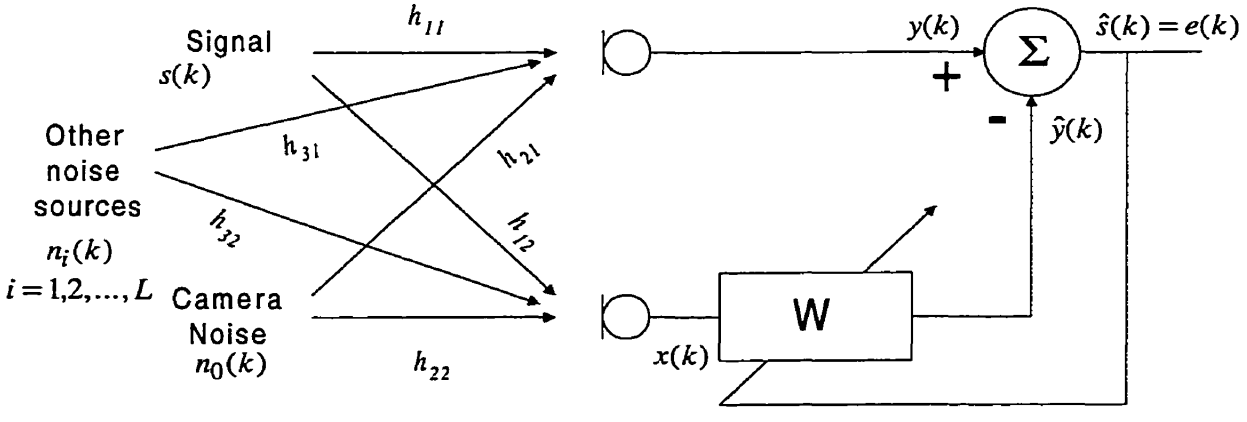


Figure 4.5 Block diagram of an ANC system under realistic conditions.

It can be seen that a new path h_{12} has been added. This path represents the leakage of the desired signal into the reference input $x(k)$. This leakage places an upper bound on

the amount of noise cancellation that can be achieved. The leakage also causes the noise-reduced signal at the ANC output to be somewhat distorted. In their seminal paper on adaptive noise cancellation, Widrow *et al.* [7] showed that the maximum attainable signal-to-noise ratio at the ANC output is equal to the reciprocal (at all frequencies) of the signal-to-noise ratio of the reference input,

$$\max\{SNR_{output}\} \leq \frac{1}{SNR_{reference}} . \tag{4.40}$$

Therefore, as an example, if the level of the signal is 20 dB below the level of the noise in the reference input, then a maximum of 20 dB of noise reduction can be achieved by this ANC system.

Widrow *et al.* also derived an expression to estimate the amount of signal distortion D at the ANC output,

$$D \cong \frac{SNR_{reference}}{SNR_{primary}} , \tag{4.41}$$

where $y(k)$ is the primary signal and $x(k)$ is the reference signal.

Equation (4.41) shows that the signal distortion will be higher if the signal-to-noise ratio at the reference input is high and the signal-to-noise ratio at the primary input is low. Of course, if there is no signal leakage into the reference input, then the output signal will not be distorted.

Referring back to Figure 4.5, it is frequently assumed that the path h_{22} from the noise source to the reference input is equal to 1. However, in some situations, this assumption is not valid. This is certainly the case for the camera noise problem, and thus the implications of $h_{22} \neq 1$ must be considered. To understand the potential impact of this, we examine the operation of the ANC system in the z-domain. The error is

$$Y(z) - \hat{Y}(z) = S(z)H_{11}(z) + N(z)H_{21}(z) - N(z)H_{22}(z)W(z) \tag{4.42}$$

The error is minimized if

$$W(z) = \frac{H_{21}(z)}{H_{22}(z)} . \tag{4.43}$$

Therefore, in general $W_{opt}(z)$ must include the inverse of $H_{22}(z)$. If $H_{22}(z)$ is non-minimum phase, then $H_{22}^{-1}(z)$ will be unstable and as a result, the adaptive filter will not converge to the value given in (4.43). It is well established that acoustic impulse responses in rooms are generally non-minimum phase [64,65] and thus, in the case of reducing camera noise, one may be faced with a non-minimum phase h_{22}. It is therefore

possible that the amount of noise reduction attainable in the camera noise application may be limited by this factor.

The more realistic ANC scenario depicted in Figure 4.5 also introduces the existence of additional noise sources, n_i; i=1,2,...L. To examine the effect of these additional noise sources, we begin by defining the coherence function between the primary signal $y(k)$ and the reference signal $x(k)$ as

$$\gamma_{yx}(\omega) = \frac{S_{yx}(\omega)}{\sqrt{S_{yy}(\omega)\, S_{xx}(\omega)}} \tag{4.44}$$

where ω denotes the frequency of interest. $S_{yx}(\omega)$ is the complex cross-power spectrum

$$S_{yx}(\omega) = \sum_{l=-\infty}^{\infty} r_{yx}(l) e^{-j\omega l} \tag{4.45}$$

where

$$r_{yx}(l) = E[y(k)x(k-l)] \tag{4.46}$$

is the cross-correlation. Both $y(k)$ and $x(k)$ are assumed to be wide-sense stationary random processes. $S_{yy}(\omega)$ and $S_{xx}(\omega)$ are the power spectra of $y(k)$ and $x(k)$ respectively. For convenience, we define the magnitude-squared coherence,

$$C_{yx}(\omega) \equiv |\gamma_{yx}(\omega)|^2 = \frac{|S_{yx}(\omega)|^2}{S_{yy}(\omega) S_{xx}(\omega)} . \tag{4.47}$$

Consider the frequency-domain representation of the cost function as defined in (4.3) and (4.7).

$$S_{ee}(\omega) = E[|Y(\omega) - \hat{Y}(\omega)|^2] \tag{4.48a}$$

$$= E[|Y(\omega) - W(\omega)X(\omega)|^2] \tag{4.48b}$$

$$= S_{yy}(\omega) - W^*(\omega)S_{yx}(\omega) - W(\omega)S_{yx}^*(\omega) + |W(\omega)|^2 S_{xx}(\omega) \tag{4.48c}$$

Completing the squares and substituting the definition for magnitude-squared coherence in (4.47), gives

$$S_{ee}(\omega) = [1 - C_{yx}(\omega)]S_{yy}(\omega) + \left| W(\omega) - \frac{S_{yx}(\omega)}{S_{xx}(\omega)} \right|^2 S_{xx}(\omega). \tag{4.49}$$

The above equation is minimized when the filter $W(\omega)$ is equal to

$$W_{opt}(\omega) = \frac{S_{yx}(\omega)}{S_{xx}(\omega)} \tag{4.50}$$

which is in agreement with the normal equation found in (4.13). Assuming this optimum solution, (4.49) reduces to

$$S_{ee}(\omega)=[1-C_{yx}(\omega)]S_{yy}(\omega)\ . \tag{4.51}$$

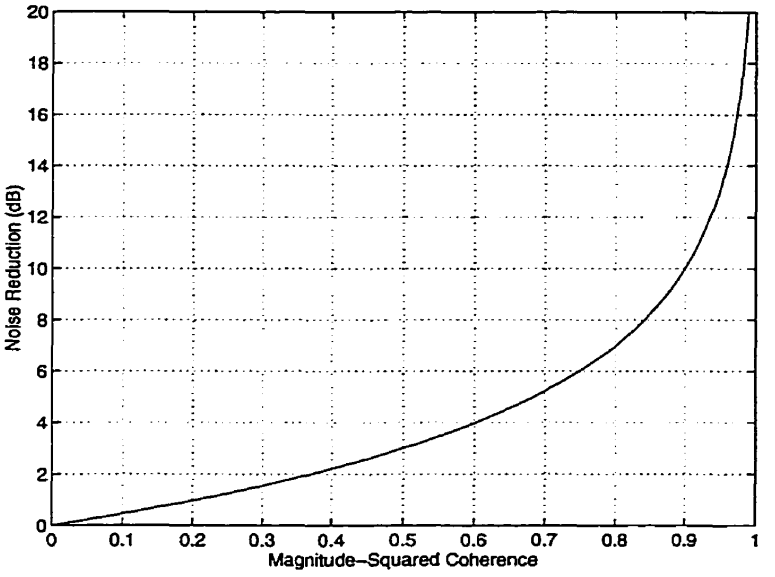


Figure 4.6 Relation between magnitude-squared coherence and maximum noise reduction attainable.

Therefore, the performance of the ANC system is dependent upon the coherence between $y(k)$ and $x(k)$. A high coherence ($C_{yx}(\omega) \Rightarrow 1$) implies a small residual error. The maximum noise reduction attainable by the ANC system at frequency ω is given by $-10\log_{10}(1-C_{yx}(\omega))$ and is plotted in Figure 4.6. It can be seen from the figure that a magnitude-squared coherence of at least 0.99 is required in order to obtain 20 dB of noise reduction.

It can be shown that the magnitude-squared coherence $C_{yx}(\omega)$ is independent of the acoustic paths h_{11}, h_{12}, h_{21}, and h_{22}. The magnitude-squared coherence is lowered by the presence of additional independent noise sources in the ANC system as depicted by n_i; i=1,2,...,L in Figure 4.5. The additional noise sources could be the result of the background acoustic noise of the room, or electronic noise in the microphones and preamplifiers. Since the background noise of the room is often out of the control of the user, it may be a limiting factor in the amount of noise reduction attainable.

The additional noise sources may also be the result of a distributed noise source. As described earlier, a camera is composed of many components each contributing to the overall noise output of the camera. Moreover, the various noise sources are physically distributed in space and they have separate acoustic paths to the reference microphone

and thus reduce the magnitude-squared coherence. That is, the camera noise at the reference microphone is

$$x(k) = \sum_{i=0}^{P} n_i(k) * h_{22_i} , \tag{4.52}$$

where n_i represents the components of the distributed noise source, and h_{22_i} represents the paths from the noise components to the reference input. The effects of a distributed noise source have been studied in other ANC applications.

Early studies by Boll and Pulsipher [63,66] examining the potential performance of acoustic ANC systems predicted that 10 to 20 dB of noise reduction was achievable. In these studies, there was no leakage of the desired speech signal into the reference microphone and the two microphones were placed several meters apart. An LMS adaptive filter having 1500 taps was used and took approximately 15 s to converge fully. This relatively long adaptation time was the result of choosing a small enough value of μ to not create an audible echo as described earlier. Following on these results, Harrison *et al.* [67,68] used ANC to reduce the noise level in the voice communications system imbedded in the oxygen facemasks worn by fighter aircraft pilots. Here, the primary microphone was placed inside the facemask while the reference microphone was placed about 6 cm away on the outside of the mask. Harrison *et al.* reported an average reduction in the noise level of about 11 dB in their simulations.

At about the same time, Darlington, Wheeler, and Powell [69,70] also examined the use of ANC for the cockpit noise problem. They found however, that the results reported by Harrison *et al.* were overly optimistic due to the simplified simulation used in that study. Darlington, Wheeler, and Powell found that, due to the distributed nature of the noise source(s) in a cockpit, acoustic ANC was only effective at frequencies below about 1 kHz. Based on work by Piersol [46], they showed that the distributed noise source(s) in the cockpit lowered the coherence between the two input signals. Specifically, Piersol showed that in a diffuse (spherically isotropic) noise field, the magnitude-squared coherence between two omnidirectional receivers will be

$$C_{yx}(\omega)=|\gamma_{yx}(\omega)|^2=\left[\frac{\sin(\omega d/c)}{\omega d/c}\right]^2 \tag{4.53}$$

where d is the distance between the two receivers, and c is the speed of sound. A distributed noise source in a reverberant environment will behave as a diffuse noise field.

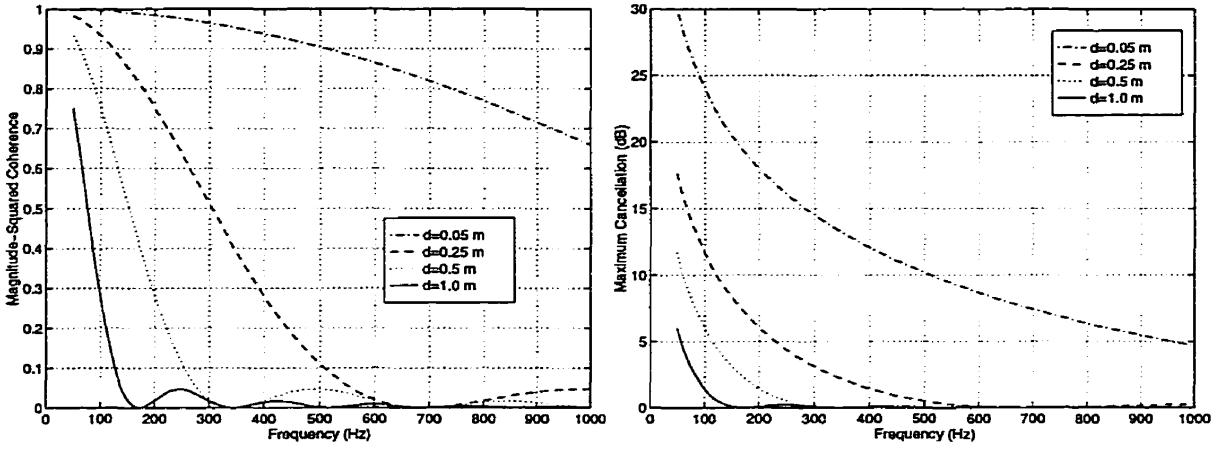


Figure 4.7 (a) Magnitude-squared coherence and (b) maximum theoretical cancellation versus frequency as a function of distance between receivers in a diffuse noise field.

As shown earlier, the coherence determines the maximum attainable degree of noise reduction. Equation (4.53) indicates that the coherence is inversely related to the distance between the receivers, and thus to obtain a high degree of noise cancellation at the higher frequencies, the primary and reference microphones must be placed close together. This contradicts the typical ANC approach wherein the two receivers are spaced far apart to avoid leakage of the desired signal into the reference input. Equation (4.53) also indicates that, for a given distance, the coherence will decrease with increasing frequency as shown in Figure 4.7a.

A study by Rodriguez *et al.* [71] confirmed these findings and concluded that ANC is ineffective in the presence of a distributed noise source. More recently, Elko examined the possibility of using first-order differential microphones to increase the coherence and thus improve the performance of an ANC system in a spherically isotropic noise field [33,72,73]. He found that the use of directional microphones did not significantly increase the coherence, and so no improvement in the amount of attainable noise reduction can be expected. In fact, depending on their orientation relative to the sources, directional microphones may reduce the coherence.

Elko also points out that if the length of the adaptive filter is too short relative the reverberation time of the room, the reverberant energy which is not accounted for by the adaptive FIR filter, will act like a spherically isotropic noise field. This result can also be predicted from Piersol's work.

In the present application, the camera acts as a distributed noise source and thus, in a reverberant environment, it will behave as a diffuse noise field. Furthermore, due to the relatively high sampling rate required, and the potentially long reverberation times, the

length of the adaptive filter $\mathbf{w}(k)$ will necessarily account for only a portion of the acoustic impulse response, h_{21}. Therefore, the camera noise convolved with the remaining portion of the impulse response will appear as a spherically isotropic noise field. These two factors will conspire to lower the magnitude-squared coherence and thus severely limit the performance of the ANC system in the camera noise application.

4.3.4 Summary of ANC methods

In this section, noise cancellation methods were introduced in the context of the Wiener filter. An adaptive noise cancellation scheme was derived based on the well known LMS algorithm. While the LMS algorithm is attractive from the point of view of computational complexity, its rate of convergence is relatively slow. Transform domain variants of the LMS algorithm can significantly improve the rate of convergence with only a small increase in computational complexity.

The fundamental limitations of the ANC method were examined. It was shown that leakage of the desired signal into the reference input imposed a bound on the amount of noise reduction attainable, and also caused the output signal to be distorted. The performance of the ANC system could also be limited if the path h_{22} was non-minimum phase. Finally, the relation between the inter-channel coherence and the ANC's performance was derived. Several factors can reduce the coherence including a distributed noise source, and an insufficiently long adaptive FIR filter.

Extensions to the basic ANC method have been proposed in the literature to address many of its limitations. However, in its basic form, ANC has inherent limitations which constrain its usefulness in many practical situations. Even under the best of conditions, acoustic ANC systems do not achieve more than about 20 dB of broadband noise reduction. And, while this amount of noise reduction is impressive, it is not sufficient for reducing higher levels of camera noise. Therefore, we must consider other methods of resolving the camera noise problem.

4.4 Blind Signal Separation

In recent years there has been a great deal of research into the problem of blind signal separation. Stated simply, blind signal separation consists of separating n signals which have been mixed together in some unknown manner. More specifically, given n sources which have been mixed together in some way and recorded at n receivers, the goal is to recover the n original signals. The problem is *blind* in the sense that it is assumed that

nothing is known about the mixing parameters. The only assumption is that the n source signals are mutually independent. Given this assumption, the various blind signal separation techniques take the mixture of n signals and strive to generate n independent output signals. Blind signal separation is often likened to the so-called *cocktail party effect* wherein a listener is able to focus his attention on a given source signal in the presence of other interfering sources [74].

Clearly, if successful, blind signal separation techniques would be of great benefit in many applications. In the present application the blind signal separation problem can be viewed as a generalization of the ANC problem. In the blind signal separation approach both the desired and interfering source signals of the ANC system are viewed as desired output signals. By generalizing the problem, the blind signal separation methods can overcome some of the limitations of ANC described in the previous section.

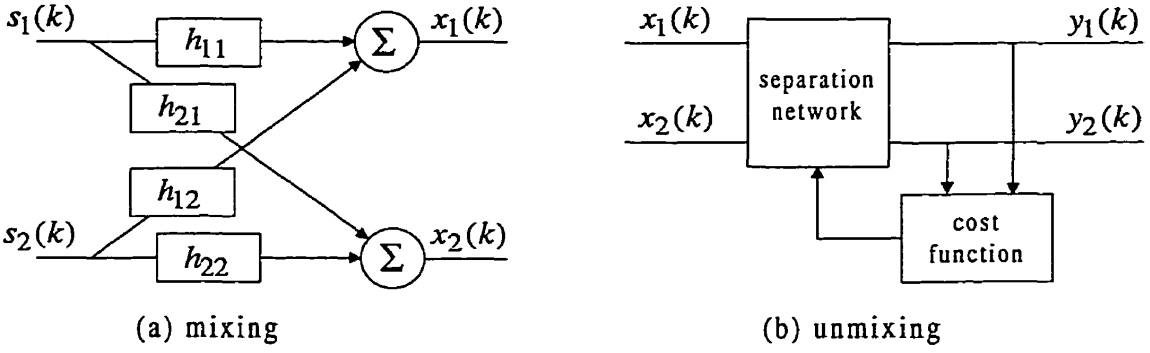


Figure 4.8 Illustration of signal separation problem for the 2×2 case: (a) the mixing paths, (b) unmixing filters.

Many researchers have limited their examination of the blind signal separation problem to the 2×2 case as depicted in Figure 4.8. The mixing portion of the blind signal separation process is shown in Figure 4.8a. Here two signals (eg. 2 talkers, or a talker and a noise source) are mixed together to form the received signals $x_1(k)$ and $x_2(k)$. More specifically, the source signals $s_1(k)$, $s_2(k)$ are mixed according to the relations described by h_{11}, h_{12}, h_{21}, and h_{22} and are collected at two receivers to form $x_1(k)$ and $x_2(k)$. Some researchers assume the simple case where h_{ij}; j=1,2 are scalars, but in general we are more interested in the convolutive mixing problem. In this case, the paths h_{ij}; j=1,2 are assumed to be FIR filters. Figure 4.8b shows the unmixing or separation algorithm where, given the mixed inputs $x_1(k)$ and $x_2(k)$, we attempt to estimate the original source signals $s_1(k)$ and $s_2(k)$.

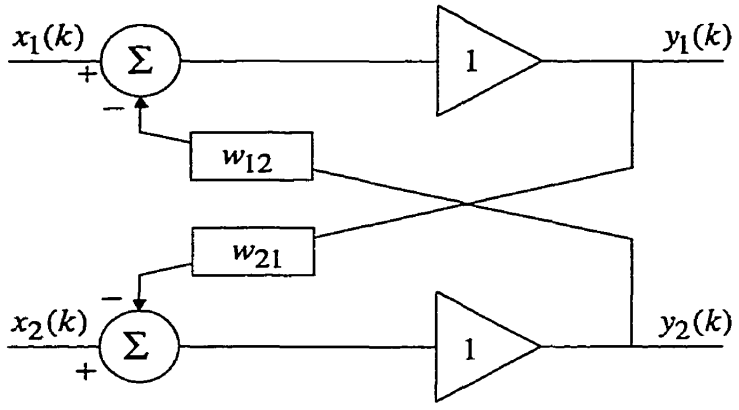


Figure 4.9 Herault-Jutten method for blind signal separation.

While some notable early work was conducted on the problem of blind signal separation, interest in this topic appears to have grown significantly with the publication of the work by Jutten and Herault [75] in 1991. They proposed a recurrent neural network approach for separating scalar mixtures as shown for the 2×2 case in Figure 4.9.

The approach can be described in matrix form,

$$\mathbf{y}(k) = \mathbf{x}(k) - \mathbf{W}(k)\mathbf{y}(k) \tag{4.54}$$

$$\mathbf{x}(k) = \mathbf{H}\mathbf{s}(k) \tag{4.55}$$

where $\mathbf{H}$ is the unknown mixing matrix. Hence,

$$\mathbf{y}(k) = [\mathbf{I} + \mathbf{W}(k)]^{-1}\mathbf{x}(k) \tag{4.56}$$

where

$$\mathbf{W} = \begin{bmatrix} 0 & w_{12} \\ w_{21} & 0 \end{bmatrix}. \tag{4.57}$$

Based on the independence criteria, Jutten and Herault adapted the elements of $\mathbf{W}$ using the simple adaptive learning algorithm

$$\frac{dw_{ij}(k)}{dk} = \mu \cdot f(y_i(k)) \cdot g(y_j(k)) \tag{4.58}$$

where μ is the adaptation parameter and f(·) and g(·) are two different odd non-linear functions such as $f(y)=y^3$ and $g(y)=\tanh(10y)$.

The Herault-Jutten algorithm has received much attention and work is ongoing to overcome its limitations and improve its performance. Nomura *et al.* [76] proposed an extension to the Herault-Jutten network to provide for delayed source signals. Cichocki *et al.* [77] found that the Herault-Jutten algorithm performed poorly when the input sig-

nals were badly scaled (i.e., weak signals mixed with strong signals) and proposed a modification which addresses this issue.

Blind signal separation using higher-order statistics (HOS) has also been studied. The appropriateness of using HOS lies in the fact that statistical independence is a much stronger property than uncorrelatedness. By using higher-order moments to test for independence of the output signals, it is possible to estimate the elements of the mixing matrix **H**.

Cordosa [78] proposed a blind identification scheme based on fourth-order normalized moments. Thi and Jutten [79] addressed the problem of separating convolutive mixtures of source signals using fourth-order cumulants, while recently Shamsunder and Giannakis [80] developed bispectrum and trispectrum based algorithms for the multichannel blind signal separation problem.

While the use of HOS for blind signal separation is attractive from a theoretical point of view, they suffer from several disadvantages which often make them impractical in real-world applications. First, calculation of higher-order cumulants and polyspectra is very computationally demanding and often requires large amounts of computer storage (memory). Secondly, reliable estimates of higher-order statistics require long data samples, making their use very difficult in situations involving non-stationary signals or a time-varying mixing process.

Bell and Sejnowski [81] proposed a technique for blind signal separation based on maximizing the entropy of non-linearly transformed output signals. The non-linearity was obtained through the squashing function tanh(·). They reported very good results including the "nearly perfect" separation of up to 10 digitally mixed speech signals. Bell and Sejnowski's algorithm was also applied to the problem of blind deconvolution.

Torkola [82,83] addressed two limitations of the Bell and Sejnowski algorithm. First, he extended the algorithm to the separation of delayed signals. Torkola then showed how the entropy maximization method could be used to address the problem of convolutive signal mixtures. Most recently, Lee *et al.* [84] extended Torkola's work to account for non-minimum phase mixing of the source signals. They tested their algorithm by attempting to separate 2 talkers in a real room. They report very good performance and conclude that it should now be possible to apply the algorithm to real-world blind signal separation problems.

Smaragdis proposed a frequency-domain extension to the work of Torkola and Lee *et al.* [85]. His intent was to reduce the statistical dependence between the taps of the separation filters while also reducing the computational complexity. This is akin to the use of transform domain variants of the LMS algorithm to increase the rate of convergence. Separation of artificially mixed signals was demonstrated to be quite good, however performance of the algorithm in a real-world situation remains unproven.

The final approach to the blind signal separation problem which we will consider consists of adaptively decorrelating the output signals. These methods only exploit second-order statistics and do not take advantage of the information contained in the higher moments. Therefore, in theory they will not perform as well as when higher order statistics are considered. This potential reduction in performance is offset by the significant reduction in computational complexity and the improved reliability of estimating the required signal statistics. Blind signal separation methods based on output decorrelation are of particular interest in the present study because of their direct relation to the ANC methods described earlier.

Algorithms based on output decorrelation for separating two signals from a scalar mixture have been developed by Canagarajah [86] as well by Van Gerven and Van Compernolle [87]. These methods however do not extend to the case of convolutive mixtures and hence, are of little interest in the present application.

The basic 2×2 blind signal separation system based on output decorrelation is shown in Figure 4.10. The input signals x_i; i=1,2 are assumed to be a result of the mixing process depicted in Figure 4.8a. In matrix form, in the frequency domain, we have

$$\mathbf{x}(\omega) = \mathbf{H}(\omega)\mathbf{s}(\omega) \tag{4.59}$$

where

$$\begin{pmatrix} X_1(\omega) \\ X_2(\omega) \end{pmatrix} = \begin{pmatrix} H_{11}(\omega) & H_{12}(\omega) \\ H_{21}(\omega) & H_{22}(\omega) \end{pmatrix} \begin{pmatrix} S_1(w) \\ S_2(w) \end{pmatrix} \tag{4.60}$$

To find $\mathbf{s}(\omega)$ from $\mathbf{x}(\omega)$ we require that $\mathbf{W}(\omega)$ be the inverse of $\mathbf{H}(\omega)$. That is, given

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega) \tag{4.61}$$

where

$$\begin{pmatrix} Y_1(\omega) \\ Y_2(\omega) \end{pmatrix} = \begin{pmatrix} W_{11}(\omega) & W_{12}(\omega) \\ W_{21}(\omega) & W_{22}(\omega) \end{pmatrix} \begin{pmatrix} X_1(w) \\ X_2(w) \end{pmatrix} \tag{4.62}$$

or equivalently

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{H}(\omega)\mathbf{s}(\omega). \tag{4.63}$$

Therefore, if

$$\mathbf{W}_{opt}(\omega) = \mathbf{H}^{-1}(\omega) \tag{4.64}$$

then we have

$$\mathbf{y}(\omega) = \mathbf{s}(\omega). \tag{4.65}$$

Thus we require

$$\mathbf{W}(\omega)=\mathbf{H}^{-1}(\omega)=\frac{1}{H_{11}(\omega)H_{22}(\omega)-H_{12}(\omega)H_{21}(\omega)}\begin{pmatrix} H_{11}(\omega) & -H_{12}(\omega) \\ -H_{21}(\omega) & H_{22}(\omega) \end{pmatrix}. \tag{4.66}$$

Typically, $H_{11}(\omega)$ and $H_{22}(\omega)$ are assumed to be equal to 1. The task is then to find $\mathbf{W}(\omega)$ using the decorrelation criteria,

$$E[y_1(k)y_2(k+l)] = 0 \quad \forall l. \tag{4.67}$$

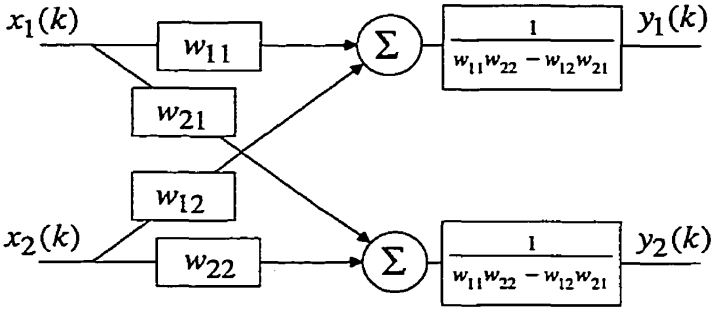


Figure 4.10 Basic 2×2 signal separation system based on output decorrelation.

In 1993, Weinstein *et al.* [88] proposed a method for blind signal separation of convolutive signal mixtures. They derived a recursive algorithm using the cross-correlations between the input and output signals of their system. Their cost function was related to the cross-correlation function [89]

$$r_{y_1y_2}(k) = E[y_1(k)y_2(k+l)] \tag{4.68}$$

$$= E[y_1(k)(x_2(k+l)+\mathbf{w}_{21}^T\mathbf{x}_1(k+l))] \tag{4.69}$$

$$= E[y_1(k)x_2(k+l)]+E[\mathbf{w}_{21}^T\mathbf{x}_1(k+l)y_1(k)]. \tag{4.70}$$

Note that the definition of $r_{y_1y_2}(k)$ includes the output signal in its formulation.

In computer simulations using low order FIR mixing filters, they obtained an increase in the signal-to-noise ratio of about 10 dB. An interesting aspect of the Weinstein *et al.*

method is that both the LMS and recursive least squares ANC systems are special cases of their blind signal separation system.

Yellin and Weinstein [90] extended the method of Weinstein *et al.* by incorporating higher-order cumulants into the criteria for determining the optimal separation filters. The higher-order statistics take advantage of the independence assumption of the two input signals. They tested their method by attempting to separate two source signals (speech and music) in a real room. Using second and forth-order statistics, they report very good results, and thus, like Lee *et al.*'s entropy maximization scheme, this method appears promising for real-world applications.

A blind signal separation algorithm was proposed by Chan *et al.* [89,91] which is similar in many ways to the method developed by Weinstein *et al.* Their method also uses an iterative time-domain algorithm and can be readily applied to the $n \times n$ case.

Chan *et al.* used a different cost function than Weinstein *et al.* Their cost function was related to the cross-correlation function

$$r_{y_1 y_2}(k) = E[y_1(k) y_2(k+l)] \tag{4.71}$$

$$= E[(x_1(k) + \mathbf{w}_{12}^T \mathbf{x}_2(k))(x_2(k+l) + \mathbf{w}_{21}^T \mathbf{x}_1(k+l))] \tag{4.72}$$

Note that this definition of $r_{y_1 y_2}(k)$ only uses the input signals x_i; i=1,2 and not the output signals.

Chan *et al.* claim several advantages over the Weinstein algorithm. First, whereas the Weinstein method requires the number of lags used to calculate the cross-correlations to be equal to the length of the filters, $\boldsymbol{w}_{ij}$, the Chan algorithm does not. Therefore, the Chan method offers a degree of flexibility which may be important when using low order separation filters. Secondly, in the Weinstein method, interim values of the cross-correlations between the input and output signals must be calculated when iterating towards the optimal solutions for $\boldsymbol{w}_{ij}$. In Chan's method, the input signal correlations are used and only need to be calculated once. Therefore, the algorithm offers a significant reduction in computational complexity. A more complete comparison of these two algorithms can be found in [89]. Chan demonstrated very good performance of his algorithm for computer simulations as well as for simulations in an anechoic chamber. Unfortunately, tests of the algorithm in a real-world situation were not provided.

Molgedey and Schuster [92] proposed a method for separating signals from a scalar mixture using time delayed correlations to reduce the task of determining the mixing co-

efficients to an eigenvalue problem. More recently, Ehlers and Schuster [93] extended this work to the blind separation of convolutive mixtures and proposed a Monte Carlo approach for minimizing their cost function. They applied their algorithm to the problem of automatic speech recognition and report impressive improvements in the recognition error rate.

4.5 Comparison of the performance of ANC and BSS systems

As stated earlier, there are many similarities between blind signal separation and adaptive noise cancellation. This is particularly true for blind signal separation algorithms based on second-order statistics (i.e., output decorrelation). It was seen earlier that the leakage of the desired signal into the reference input of an ANC system will directly limit the amount of noise reduction possible. Blind signal separation systems overcome this problem by adding the separation filter w_{21} (as shown in Figure 4.10) to account for the leakage path. Therefore, blind signal separation systems have a potential advantage over ANC systems in applications where leakage is a problem.

Chan compared blind signal separation and ANC performance is the presence of signal leakage and found that, as expected, the ANC system performed poorly when there was a significant amount of leakage. Conversely, the performance of the blind signal separation system was relatively independent of the level of leakage. Oddly, Chan's results indicate that blind signal separation always performs better (by more than 10 dB) than ANC even in the absence of leakage. Moreover, his results indicate that, when the level of the signal leakage is below some value, both the ANC and blind signal separation systems actually reduce the signal-to-noise ratio from input to output. Chan's findings are counter-intuitive and contradict previous research, and should be explored further.

Another limitation of ANC systems occurs when the path h_{22} is non-minimum phase. This restricts the ability of the adaptive filter to converge. In the blind signal separation approach this problem is resolved by the filters w_{11} and w_{22}. However, most blind signal separation algorithms simply set w_{11} and w_{22} equal to 1 since additional constraints are required in order to solve for the four filters w_{ij}; i,j=1,2. Chan *et al.* offer several possible constraining conditions [89,91,94].

It was shown earlier how the performance of an ANC system is dependent upon the magnitude-squared coherence of the input signals. One way in which the coherence can be lowered is by the presence of additional noise sources. The effect of additional sources (i.e., more sources than receivers) on the performance of blind signal separation

systems does not appear to be fully addressed in the literature. It would seem that the effect of additional sources on blind signal separation systems should be similar to the effect on ANC systems, and so the blind signal separation approach should not offer any advantage in this regard. As noted earlier, a distributed noise source is equivalent to having additional noise sources.

It was shown that one way to increase the coherence in a diffuse noise field is to place the receivers closer together. It is interesting to speculate whether this would improve the performance of a blind signal separation system. Recall, that the cost function of a 2×2 blind signal separation system based on output decorrelation is related to

$$r_{y_1y_2}(k) = E[(x_1(k) + \mathbf{w}_{12}^T \mathbf{x}_2(k))(x_2(k+l) + \mathbf{w}_{21}^T \mathbf{x}_1(k+l))] \ . \tag{4.73}$$

In a diffuse sound field, the input cross-correlation terms will be affected by the distance between the receivers (see Figure 4.7). At one extreme, if the receivers are spaced very far apart, then the cross-correlation function will tend towards zero and the blind signal separation system will minimize the cost function by setting w_{21}=w_{12}=0 (i.e., $\mathbf{W}$ becomes the identity matrix). In this case, the blind signal separation system will do nothing, and the output signals will be equal to the input signals,

$$as \ \ r_{x_1x_2}(k) \Rightarrow 0, \quad \begin{cases} y_1(k) \Rightarrow x_1(k) \\ y_2(k) \Rightarrow x_2(k) \end{cases} . \tag{4.74}$$

At the other extreme, if the distance d between the receivers is very small, the impulse responses from a given source to the two receivers will begin to look similar,

$$as \ \ d \Rightarrow 0, \quad \begin{cases} h_{11} \approx h_{21} \\ h_{22} \approx h_{12} \\ x_1(k) \approx x_2(k) \end{cases} \tag{4.75}$$

and as a result, in order to minimize the cost function, the blind signal separation system will tend towards

$$as \ d \Rightarrow 0, \quad \begin{cases} w_{21} \Rightarrow 1 \\ w_{12} \Rightarrow 1 \\ y_1(k) \Rightarrow 0 \\ y_2(k) \Rightarrow 0 \end{cases} . \tag{4.76}$$

Therefore, under these conditions, the output signals will go to zero and signal separation will not occur. This is particularly true at lower frequencies where the wavelength is significantly larger than the distance d.

The above analysis, although somewhat heuristic, indicates that the performance of a blind signal separation system will be dependent on the spacing between the receivers for a given acoustic environment. It also suggests that, as for ANC, directional microphones may not be helpful in general for blind signal separation. Blind signal separation methods based on higher-order statistics may be less sensitive to this parameter. This matter needs further investigation including a mathematical framework within which performance tradeoffs may be determined.

Most simulations of blind signal separation systems use separation filter lengths that are at least as long as the mixing filters. It was noted earlier that the coherence between the input signals can be reduced if the separation filters are insufficiently long relative to the reverberation time of the room. Therefore, the performance reported in many simulations may be overly optimistic as compared to what can be expected in real-world applications.

Blind signal separation can be viewed as a generalized ANC technique with its greatest inherent advantage being its relative insensitivity to signal leakage. However, in the absence of signal leakage, blind signal separation systems (particularly those based on second-order statistics) are not likely to offer significantly better performance.

It should be noted that in some instances an ANC system is preferable to a blind signal separation system. Consider the 2-input situation where there are several desired signals and one interfering noise source. The goal of the noise reduction system in this situation is to produce a noise-free recording of the desired signals. This is possible with the ANC approach, because the user can "tell" the algorithm which signals are considered desirable and which signal is the noise, through appropriate placement of the receivers. If there is some leakage of one or more of the desired signals into the reference input, then it is not clear what the resulting output signals will be using blind signal separation since there are more signals than receivers. With ANC, the output will be the desired signals with some degree of noise reduction and signal distortion.

4.6 Results of Tests of ANC to Reduce Camera Noise

Sample recordings (see Chapter 3) of both the NFB camera and Imax-3D camera were processed using the ANC method to determine the amount of noise reduction that could be achieved. For these tests, a commonly employed variant of the LMS algorithm was used to adapt the filter. The Normalized LMS (NLMS) algorithm is often used in situations where the power levels of the input signals are subject to wide fluctuations such as

is found in speech signals corrupted by camera noise [9]. NLMS is described by the following update equations,

$$e(k) = y(k) - \mathbf{w}(k)\mathbf{x}^T(k) \tag{4.77}$$

and

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \mu \cdot \left[\frac{\mathbf{x}(k)e(k)}{c + \mathbf{x}^T(k)\mathbf{x}(k)} \right], \tag{4.78}$$

where c is a small positive constant which prevents division by zero if $\mathbf{x}(k)$ goes to zero.

The primary input consisted of the camera noise recorded with a microphone placed where a talker (actor) would stand with respect to the camera. The reference signal was recorded using a second microphone positioned next to the camera. Steps were taken to ensure that the primary and reference signals were correctly aligned in time. It should be noted that, since the recordings did not contain speech, there was no leakage of a desired signal into the reference input. Therefore, in these tests, we would expect a blind signal separation system to perform similarly to the ANC system.

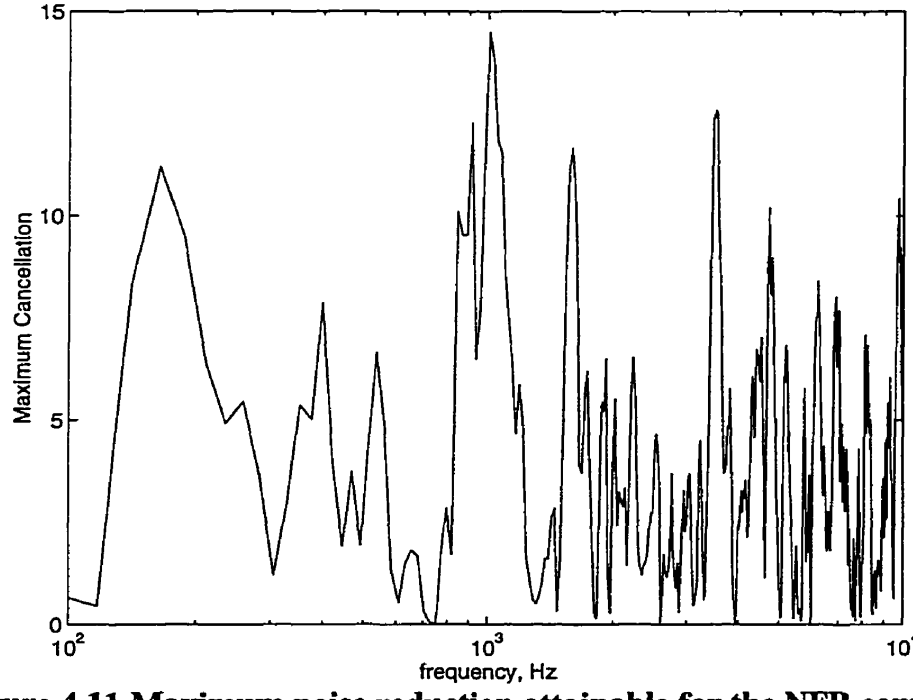


Figure 4.11 Maximum noise reduction attainable for the NFB camera.

Due to the various factors described earlier in this chapter, the ANC system only provided between about 6 and 10 dB of noise reduction for the NFB camera, and less than 6 dB for the IMAX-3D camera. These results are in good agreement with the degree of noise reduction that would be predicted from the coherence measurements shown in

Chapter 3 (see Figures 3.24 and 3.25) which showed the magnitude-squared coherence $C_{xy}(\omega)$, for some of the recordings used in the ANC tests. Recall that the magnitude-squared coherence was limited by the fact that the cameras are distributed noise sources. Using the relation described earlier in this chapter, the maximum cancellation that an ANC system can achieve can be determined from the magnitude-squared coherence. The maximum cancellation attainable for the NFB camera is plotted in Figure 4.11 and in Figure 4.12 for the IMAX-3D camera. It can be seen that, for both cameras, the cancellation is very frequency dependent, with reasonable cancellation at some frequencies, and virtually none at others. It should be recalled that the power spectrum of the camera noise is broadband and is not limited to specific frequencies (see Figures 3.19, 3.20 and 3.22).

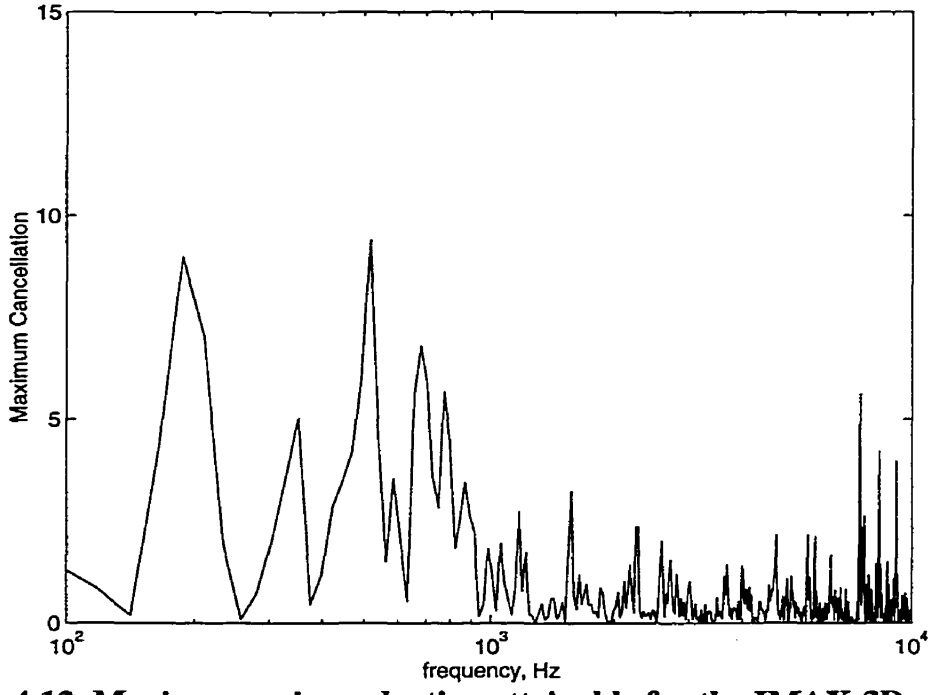


Figure 4.12 Maximum noise reduction attainable for the IMAX-3D camera.

The results of the tests indicate that the use of ANC provides no more than 10 dB of noise reduction due to the distributed nature of the camera noise. This amount of noise reduction is not sufficient for the task of removing camera noise, since a clearly audible residual noise signal will remain. Since there was no leakage of a desired signal into the reference input, it is not expected that a blind signal separation system would provide better performance than the ANC system examined here. More importantly however, this was a two-input system and thus violated a fundamental requirement for a practical camera noise reduction system.

4.7 *Adaptive Noise Cancellation Using a Synthesized Reference*

In some ANC applications, such as reducing camera noise, it is not practical or desirable to obtain a separate reference measurement of the interfering noise. In these situations it may be possible to synthesize a reference signal if the interfering noise is periodic or repetitive [9,48]. In this section, an ANC system using a synthesized reference signal for reducing the periodic component of camera noise is investigated.

It was seen in the acoustic measurements described in Chapter 3 that camera noise consists of a series of noise bursts repeating at a rate of 24 times per second. The noise bursts are seen as sharp transient peaks followed by intervals of lower level noise which extend between the peaks. A simple model of camera noise $n(k)$ was derived consisting of a periodic component $p(k)$ and a cyclical random component $c(k)$,

$$n(k) = p(k) + c(k), \tag{4.79}$$

where

$$p(k) = \sum_{l=-\infty}^{\infty} Q(k+lT). \tag{4.80}$$

with

$$Q(k) = \begin{cases} \sum_i q_i(k) & ;k=1,2,...,T \\ 0 & ;elsewhere \end{cases}, \tag{4.81}$$

where $q_i(k)$ are the components of the camera noise which repeat from frame to frame, and T is the period which is equal to the reciprocal of the film rate.

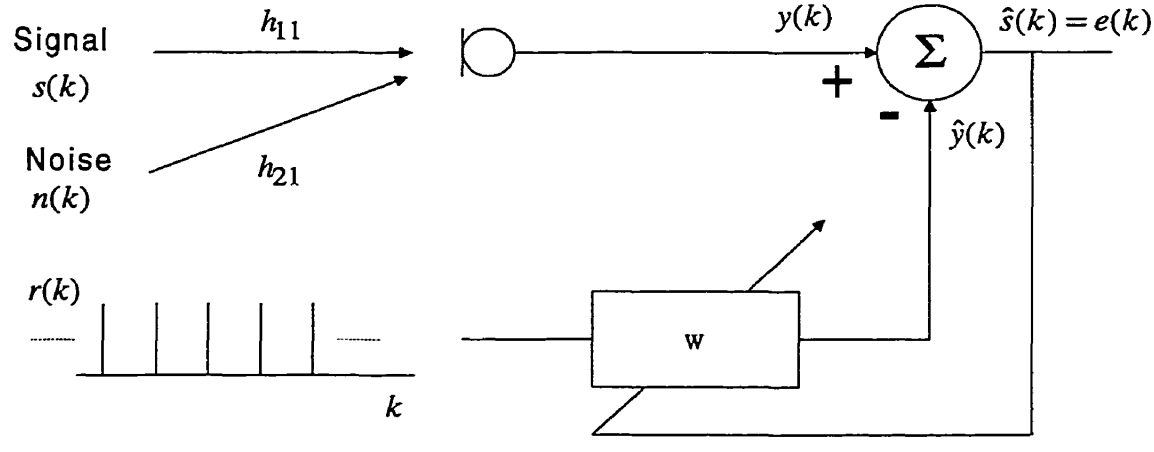


Figure 4.13 ANC system with a synthesized reference input signal.

A simple approximation to periodic component of camera noise would consist of a train of Dirac pulses occurring at a rate of 24 times per second. As shown in Figure 4.13 this approximation to camera noise was used as the reference input to the ANC system described in the previous section.

With this synthesized reference signal $r(k)$, the taps of the adaptive filter $\mathbf{w}(k)$ will converge to the periodic or "common" component $Q(k)$ of the camera noise pulses found in $y(k)$. Therefore, this approach relies on there being a strong periodic component and thus a high degree of correlation between the successive pulses of the camera noise. The periodic component, $Q(k)$ is then subtracted from the signal $y(k)$ leaving the cyclical random component $c(k)$ of the camera noise.

As a first step in evaluating the proposed technique, "ideal" camera noise was applied to the primary input of the ANC system. The ideal camera noise consisted of a single camera noise pulse (2000 samples) replicated 480 times over 20 s, and therefore $n(k)$ was equal to $p(k)$. This was done to verify that the ANC algorithm was functioning correctly, and to see whether a reduction in the periodic component of the camera noise could be achieved using a synthesized reference under best-case conditions. The results indicated that near-perfect cancellation could be achieved with this approach under these ideal conditions.

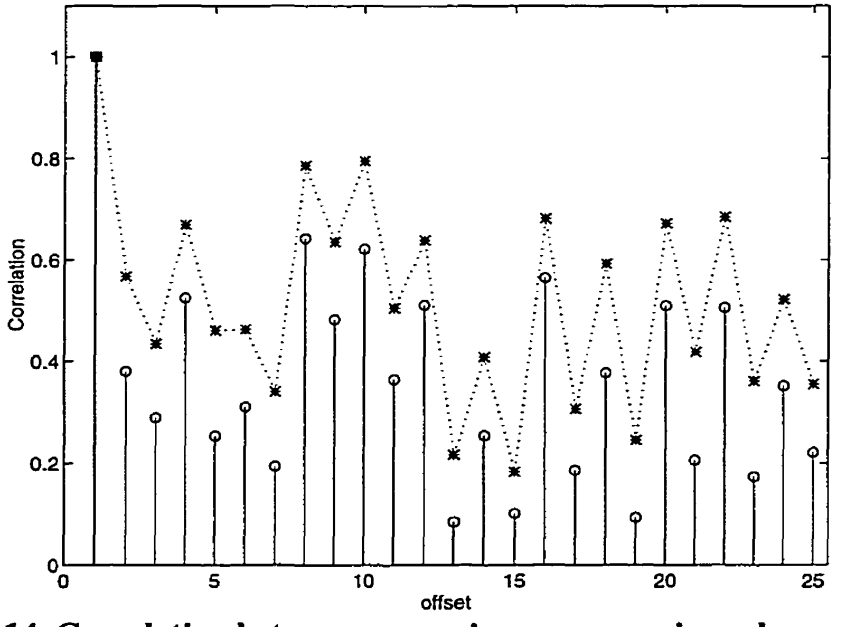


Figure 4.14 Correlation between successive camera noise pulses; open circles are non-synchronized; stars are synchronized.

The ANC system using a synthesized reference input was then implemented using a recording of real camera noise as the primary input signal. The results showed only a modest reduction in the level of the camera noise at the output of the ANC system under these conditions. Certainly, the results were not adequate for the application of removing camera noise from film soundtracks. An analysis of the adaptation process of the filter $\mathbf{w}(k)$ revealed the somewhat unexpected finding that the correlation between successive camera noise pulses was relatively low, thus suggesting that $p(k)$ is a small component of $n(k)$. This can be seen in Figure 4.14.

The open circles in the figure show the normalized correlation $\tilde{r}_{y_0y_n}(0)$ between the first (reference) pulse and each of the next 24 pulses of camera noise. As can be seen, the correlation between the reference pulse and the subsequent pulses varies significantly. Furthermore, all of the correlations are relatively low with none exceeding 0.7. These results suggest that $p(k)$ is only a small component of $n(k)$. Similar measurements on other instances of camera noise showed similar low inter-pulse correlation. The normalized inter-pulse correlation, $\tilde{r}_{y_0y_n}(0)$ is defined as,

$$\tilde{r}_{y_0y_n}(0) = \frac{r_{y_0y_n}(0)}{\sqrt{r_{y_0y_0}(0) \cdot r_{y_ny_n}(0)}} \tag{4.82}$$

where

$$r_{y_0y_n}(0) = \sum_{i=0}^{T} y_0(i) y_n(i+nT) \quad n = 1,2,3,\ldots \tag{4.83}$$

and

$$r_{y_0y_0}(0) = \sum_{i=0}^{T} y_0^2(i) \tag{4.84}$$

$$r_{y_ny_n}(0) = \sum_{i=0}^{T} y_n^2(i+nT). \tag{4.85}$$

The acoustic measurements of Chapter 3 suggested a more significant periodic component and showed that the spectral magnitude of the camera noise was relatively constant over the duration of a reel of film. Investigations were therefore conducted to determine the cause of the low inter-pulse correlation. It was hypothesized that the low correlation could be due in part to some variation (drift or jitter) in the periodic component $p(k)$ of the camera noise. Indeed, it was discovered that there are significant variations in the timings of the individual camera noise pulses. This is illustrated in Figure 4.15.

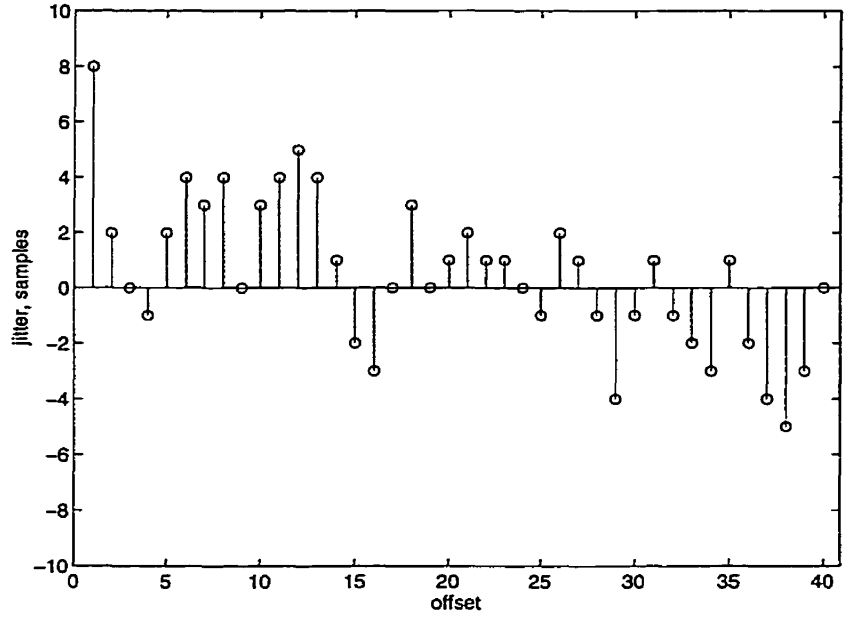


Figure 4.15 Example of jitter in the arrival times of the pulses of the camera noise.

The figure shows the jitter in the relative times of arrival (in samples) of the individual peaks of the camera noise for 40 pulses. A jitter value of 0 samples indicates that the time between noise pulses was exactly 2000 samples pulses (corresponding to the film rate of the camera at a sampling rate of 48 kHz). The points in the figure were derived by adjusting the relative time, τ between the reference pulse and each of the subsequent pulses until the maximum correlation between that pulse and the reference pulse was obtained,

$$\tau \; such \; that \; \max\{\tilde{r}_{y_0 y_n}(\tau)\} \tag{4.86}$$

where

$$\tilde{r}_{y_0 y_n}(\tau) = \frac{r_{y_0 y_n}(\tau)}{\sqrt{r_{y_0 y_0}(\tau) \cdot r_{y_n y_n}(\tau)}} \tag{4.87}$$

with

$$r_{y_0 y_n}(\tau) = \sum_{i=0}^{T-|\tau|-1} y_0(i) y_n(i+nT+\tau) \quad n = 1,2,3,\ldots \tag{4.88}$$

and

$$r_{y_n y_n}(\tau) = \sum_{i=0}^{T} y_n^2(i+nT+\tau). \tag{4.89}$$

It can be seen from the figure that there is considerable jitter in the timings of the points of maximum correlation. This implies that there is some variability in the mechanical workings of the camera which causes slight differences in the inter-pulse timing of the periodic component $Q(k)$. The jitter described in Figure 4.15 limits the performance of an ANC system based on a synthesized reference input signal.

To further investigate this matter, the ANC system was modified so that the timings of the Dirac pulses in the reference input were adjusted to account for the jitter. Specifically, the location of each individual Dirac pulse in the pulse train was synchronized to the pulses of the camera noise to obtain maximum correlation between noise pulses. Referring back to Figure 4.14, the upper curve consisting of stars joined by a dotted line indicates the maximum correlation between each pulse and the reference pulse using the ANC system with a synchronized reference input signal. It can be seen that the correlations are consistently higher for each pulse and thus this method is expected to provide better noise cancellation. However, the correlation is still relatively low for the purpose of ANC.

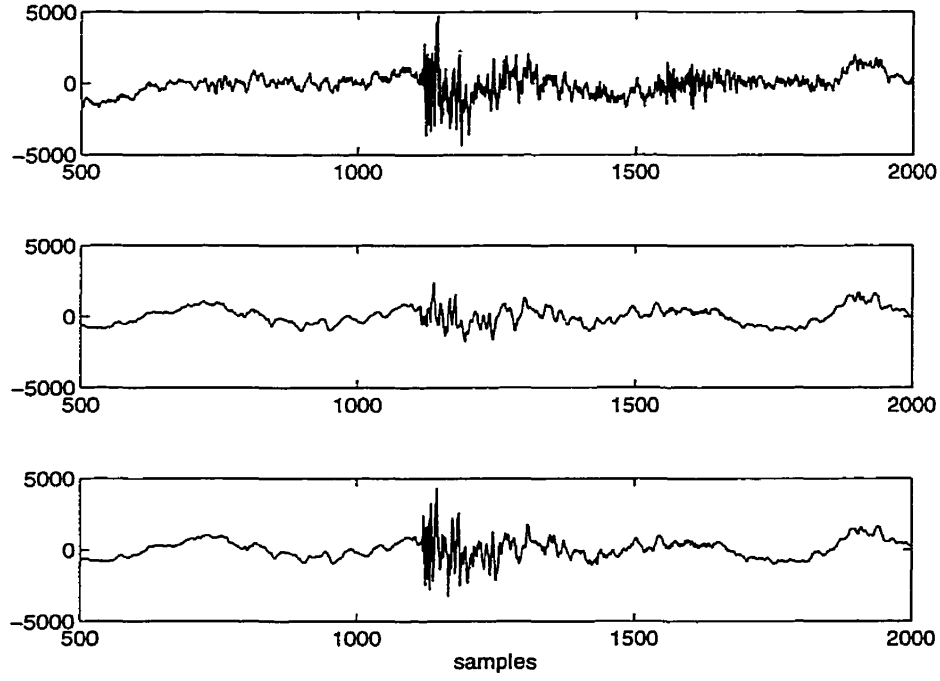


Figure 4.16 Converged values of W for the non-synchronized (middle panel) and synchronized (lower panel) ANC systems versus typical pulse (upper panel).

To compare the performance of the synchronized and non-synchronized ANC systems, it is instructive to examine the value to which the adaptive filter $\mathbf{w}(k)$ converged for each system. This comparison is provided in Figure 4.16. The upper plot shows one of

the noise pulses used in the simulation and represents a typical noise pulse. The middle plot shows the value to which the adaptive filter $\mathbf{w}(k)$ converged for the non-synchronized ANC system. It can be seen that, while the basic features of the upper plot are seen in the middle plot, the details of the pulse are not represented. That is, the periodic component $Q(k)$ obtained using the non-synchronized ANC system does not describe the detailed structure of each pulse sufficiently well. Therefore, when subtracted from the primary signal $y(k)$, the periodic component does not significantly reduce the level of the individual noise pulses.

The lower plot in Figure 4.16 shows the value to which the adaptive filter $\mathbf{w}(k)$ converged for the synchronized ANC system. It can be seen that the details (the higher frequencies) are better represented using the synchronized ANC system. Indeed, it is sensible to assume that the correlation between the high frequency components of the pulses is most affected by the jitter between noise pulses.

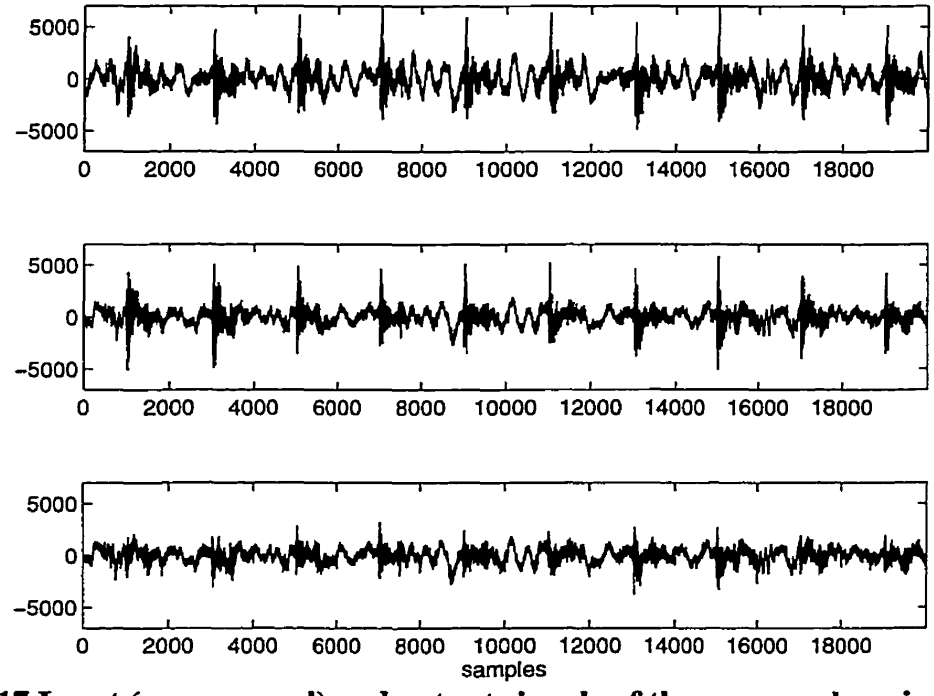


Figure 4.17 Input (upper panel) and output signals of the non-synchronized (middle panel) and synchronized (lower panel) ANC systems.

Figure 4.17 shows the input and output signals of both the synchronized and non-synchronized ANC systems. The upper plot is the input signal, the middle plot is the output of the non-synchronized ANC system, and the lower plot is the output of the synchronized ANC system. From the figure it can be seen that the non-synchronized ANC sys-

tem output has reduced the level of the noise to some extent. Specifically, the low frequency ringing seen between pulses is noticeably reduced. However, the peaks of the pulses remain largely uncanceled. In the lower plot it can be seen that the synchronized ANC system reduces both the low frequency ringing and substantially reduces the peak (transient) portion of each pulse. Therefore, by synchronizing the noise pulses for maximum correlation, the ANC system using a synthesized reference signal is providing a greater degree of noise reduction.

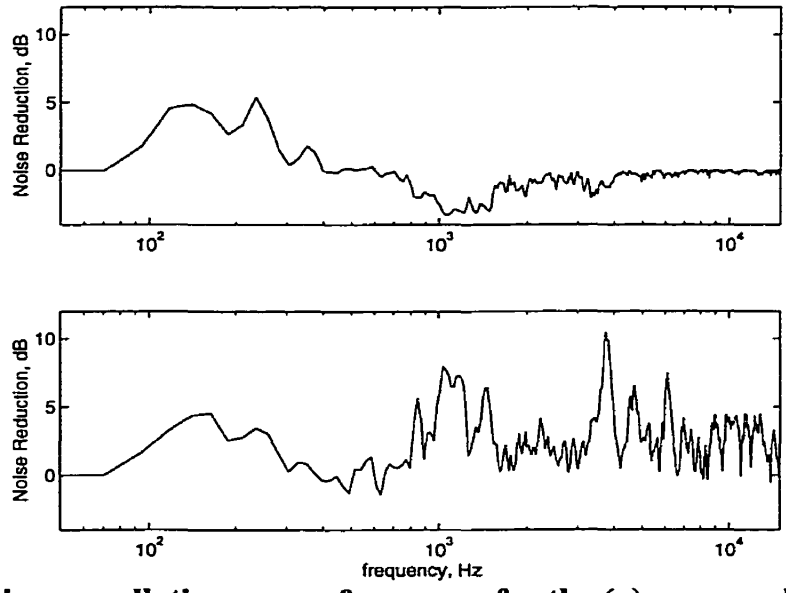


Figure 4.18 Noise cancellation versus frequency for the (a) non-synchronized and (b) synchronized ANC systems using a synthesized reference input signal.

In order to better compare the performance of the synchronized and non-synchronized ANC systems, it is instructive to examine the degree of noise reduction in the frequency domain. The upper curve in Figure 4.18 plots the noise reduction versus frequency for the non-synchronized ANC system. A reduction of about 5 dB is obtained at frequencies between 100 Hz and 300 Hz. At frequencies near 1 kHz, the non-synchronized ANC system actually increases the level of the noise in the output signal by about 1 dB. Above 1 kHz the non-synchronized ANC system does nothing.

The lower curve in Figure 4.18 plots the noise reduction obtained using the synchronized system. The performance at low frequencies is comparable to that obtained with the non-synchronized ANC system. At mid frequencies, near 1 kHz, the noise reduction is as much as 7 dB. Furthermore, the noise reduction at higher frequencies (above 2 kHz) ranges from 3 to 10 dB. Therefore, the synchronized ANC system performs significantly better than the non-synchronized ANC system. It should also be noted that the improve-

ment in performance occurs primarily at higher frequencies and so, as expected, the jitter affects primarily the higher frequencies. On average, the amount of noise reduction obtained here is about 5 dB less than the maximum attainable using traditional 2-input ANC (see Figure 4.11).

The results suggest that the model of camera noise should be modified to include jitter,

$$n(k) = p(k) + c(k), \tag{4.90}$$

$$p(k) = \sum_{l=-\infty}^{\infty} Q(k + lT + \xi(k)), \tag{4.91}$$

where $\xi(k)$ is a random variable which introduces jitter in the timing of $Q(k)$ within $p(k)$.

The results described above were for the NFB camera. To determine whether the above results can be extended to other cameras, the same analysis was repeated for two IMAX cameras.

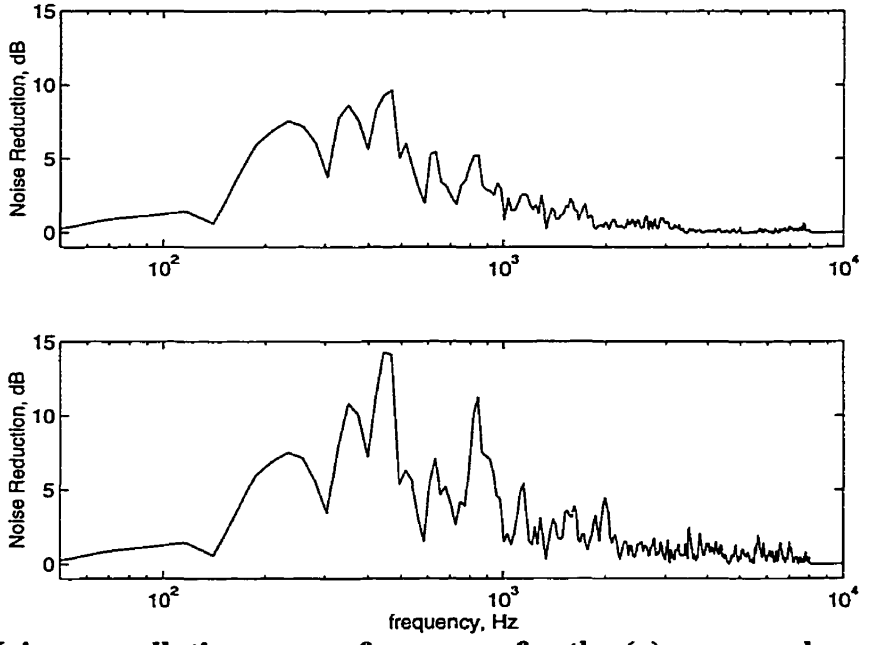


Figure 4.19 Noise cancellation versus frequency for the (a) non-synchronized and (b) synchronized ANC systems for the IMAX camera (MSM 9801).

Figure 4.19 shows the amount of noise cancellation as a function of frequency which was achieved for the standard IMAX camera. The upper curve plots the noise reduction attained using the non-synchronized ANC system. For frequencies between 200 Hz and 500 Hz a reduction of between 7 and 10 dB is obtained. About 5 dB of noise reduction is obtained for the range between 500 Hz and 1 kHz. The performance of the ANC system

falls off quickly above 1 kHz and virtually no noise reduction is obtained The lower curve in Figure 4.19 shows the noise reduction obtained using the synchronized ANC system. Except for the very low frequency range, the synchronized ANC system consistently outperforms the non-synchronized system and more than 10 dB of noise reduction is obtained at some frequencies.

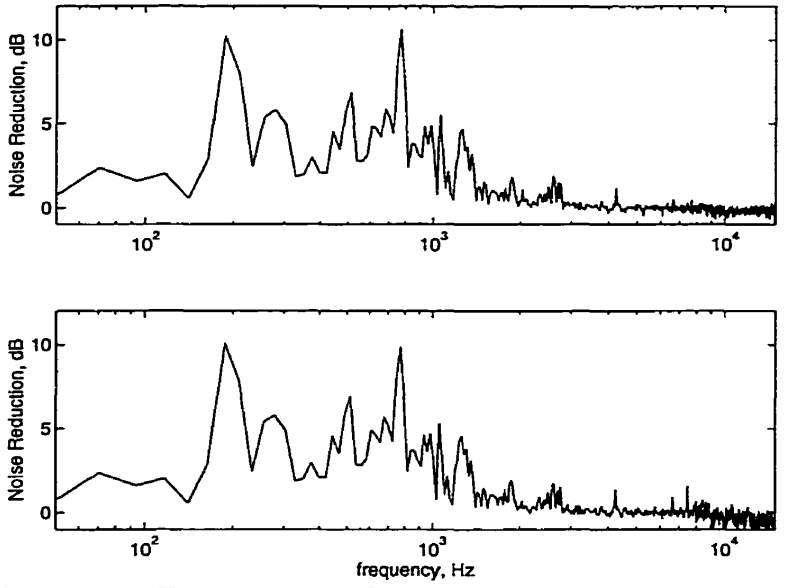


Figure 4.20 Noise cancellation versus frequency for the (a) non-synchronized and (b) synchronized ANC systems for the IMAX-3D camera.

Figure 4.20 shows the degree of noise cancellation as a function of frequency obtained for the IMAX-3D camera, with the results for the non-synchronized ANC system in the upper plot and the results for the synchronized system in the lower plot. Here, the noise reduction is limited primarily to frequencies below 2 kHz. A comparison of the two plots indicates that, for this camera, synchronizing the synthesized reference signal to the camera noise pulses does not improve the performance of the ANC system. The poor performance at higher frequencies indicates that the noise pulses are not correlated at these frequencies. This may be due to the extremely complex mechanical workings inside the IMAX-3D camera, as well as the large physical size of the camera. Interestingly, for the IMAX -3D camera, the amount of noise reduction obtained using the synthesized reference input signal is equivalent to the maximum attainable using a traditional 2-input ANC system (see Figure 4.12).

The results of this section indicate that the synthesized reference ANC system with pulse-synchronization provides almost as much noise reduction as the standard 2-input ANC system. The synthesized reference system is preferred since it is a single input ap-

proach as required for a camera noise reduction system. It should be noted, however, that the system does not perform well enough to reduce the camera noise to an acceptable level. This is because the periodic component $Q(k)$ is relatively weak, and thus the inter-pulse correlation is too low to obtain the necessary degree of noise reduction. Nonetheless, the single-input system may be of some benefit if used in conjunction with other noise reduction schemes.

4.8 Summary

In this chapter, noise reduction methods using adaptive filtering were examined. Adaptive noise cancellation based on the well know LMS algorithm was described and its limitations were identified. Transform based variants of the algorithm were seen to provide improved performance while keeping the computational complexity to a practical level. Factors which can limit the amount of noise reduction attainable with ANC were investigated including; signal leakage and low inter-channel coherence. Background noise, a distributed noise source, and the distance between microphones were all shown to contribute to low inter-channel coherence.

Blind signal separation was introduced as a generalization of ANC. Blind signal separation overcomes the signal leakage problem of ANC but is still sensitive to background noise, a distributed noise source, and microphone spacing. Tests of an ANC system to reduce camera noise achieved about 10 dB of noise reduction which is insufficient for the camera noise application.

To take advantage of the repetitive nature of camera noise, an ANC system using a synthesized reference input was proposed. It was found that, due to low inter-pulse correlation, the noise reduction obtained using this system was limited. One factor which was shown to contribute to the low inter-pulse correlation was the jitter in the arrival time of the camera noise pulses. The performance of the ANC system was improved by synchronizing the reference signal to the input signal to maximize the inter-pulse correlation. With a synchronized reference signal, the level of noise reduction was comparable to that obtained with the standard 2-input ANC system. Since a key requirement of the camera noise reduction system was that it be a single-input scheme, the synchronized synthesized reference approach is preferred. Given this finding, the mathematical model of camera noise was modified to include inter-pulse jitter.

Camera noise is composed of many sources of noise such as: the opening and closing of the shutters, the movement of the film from the supply reel to the take-up reel, the

movement of the sprockets which feed the film through the camera, the vibration of the camera enclosure, and the rotation of the motors which drive the entire mechanism. While some of these noise sources are directly related to the film rate, those which are not will tend to reduce the inter-pulse correlation and will thus limit the performance of the ANC system with a synthesized reference signal. Specifically, since the periodic component is only part of the overall camera noise, this method provides between 5 dB and 10 dB of noise reduction under best case conditions. This is insufficient for the camera noise application and so, by itself, this approach is not viable. However, the method may be somewhat beneficial if used in conjunction with other noise reduction schemes which address the non-periodic component of the camera noise.

5. SIGNAL ENHANCEMENT TECHNIQUES BASED ON ESTIMATION OF THE SHORT-TIME SPECTRAL MAGNITUDE

5.1 Introduction

In this chapter frequency domain based techniques which have been developed to reduce the level of an interfering background noise source are examined. The underlying noise reduction algorithm is commonly referred to as "spectral subtraction" and is based on estimating the short-time spectral magnitude of the signal. A key advantage of the spectral subtraction technique is that it can be applied when only the audio signal contaminated by camera noise is available. That is, unlike the adaptive noise cancellation techniques described in Chapter 4, spectral subtraction does not require direct access to the noise source in the form of a reference input. Therefore, the spectral subtraction approach is well suited to the problem of camera noise and allows the possibility of using the technique for restoring the soundtracks of older films.

In Chapter 3 it was seen that camera noise can be modeled as the sum of a periodic component and a cyclical random noise component. The adaptive noise cancellation method using a synthesized reference signal was shown in Chapter 4 to reduce the periodic component by about 10 dB. The spectral subtraction based methods examined in this chapter will be shown to successfully reduce the cyclical random component as well as the periodic component of the camera noise.

Signal enhancement based on estimating the spectral magnitude of the signal was first proposed by Weiss *et al.* in 1974 [12]. A more comprehensive study of the technique was presented four years later by Boll [13,14] who appears to have discovered the technique independently of Weiss *et al.* Boll applied the spectral subtraction technique as a preprocessor to a speech compression algorithm in a communications system. The algorithm was intended to work with both narrowband periodic noise and broadband colored noise.

The various spectral subtraction based noise reduction algorithms described in this chapter were initially developed for military applications in an effort to improve the intelligibility of speech signals under extremely adverse noise conditions. For example, variations of the algorithm have been used to enhance speech communications in heli-

copter and jet aircraft cockpits where the signal-to-noise ratio was in the range of -5 dB to +5 dB [13,14,15]. In general it was found that under these extremely adverse conditions the various spectral subtraction algorithms did not tend to provide any improvement in speech intelligibility. In fact, in some instances, the algorithms have been found to reduce the intelligibility of the speech [13,16]. The algorithms however, were used with some success as a pre-processor to data/bandwidth compression systems [95]. The spectral subtraction technique was also shown to improve the performance of digital LPC vocoders [15,27]. While the various algorithms may not improve speech intelligibility, most have been found to provide a significant improvement in the perceived quality of the resulting speech signal. It is in this context that the spectral subtraction algorithms were considered as potential means of reducing camera noise in film soundtracks.

In a film soundtrack, the signal-to-noise ratio (speech to camera noise) is expected to be significantly greater than the -5 dB to +5 dB described above (see Section 3.5.6). Furthermore, it is known that speech intelligibility is not a concern in environments having a signal-to-noise ratio greater than about +15 dB [96,97]. Therefore, the camera noise on a film soundtrack is not expected to have any effect on speech intelligibility. Rather, the camera noise may have a significant effect on the quality of the audio (usually speech) signal. Therefore, the algorithms examined in this chapter were explored solely as a means of enhancing the perceived quality of the audio signal on the film soundtrack.

The noise reduction algorithms described in this chapter assume that a noise $n(k)$ has been added to a stationary random signal $s(k)$, and that $n(k)$ and $s(k)$ are independent of each other*. It is assumed that the noise floor which is present in the short-time spectrum can be reduced by subtracting an estimate of it from the spectral magnitude of the noisy speech. Spectral subtraction noise reduction algorithms attempt to estimate the short-time spectral magnitude of the clean signal and then use the phase from the noisy signal $y(k)$, to recover an estimate $\hat{s}(k)$ of $s(k)$. The various algorithms differ primarily in the way in which the spectral magnitude of $s(k)$ is estimated. The class of noise reduction algorithms described in this chapter depend on the fact that the short-time spectral magnitude of a speech signal is perceptually important whereas its phase is relatively unimportant [13,18,19].

---

* An example of a non-independent noise is the quantization noise resulting when appropriate dithering has not been applied [S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and Dither: A Theoretical Survey," *J. Audio Eng. Soc.* vol. 40, no. 5, May 1992].

This chapter provides a detailed description of the underlying concepts of the spectral subtraction technique for signal enhancement. This is done in the context of describing the method proposed by Boll. Variations and enhancements to this underlying technique which make it particularly effective at removing camera noise from film soundtracks are described in subsequent chapters. Specifically, by decomposing the spectral subtraction process into subbands and sub-frames, the noise reduction process can be matched to the characteristics of the camera noise. Also, by incorporating a perceptual model into the spectral subtraction algorithm, the noise reduction process removes only those components of the camera noise which are audible at each instant in time. Both of these refinements to the spectral subtraction algorithm share the philosophy of reducing the amount of processing applied to the noisy signal, thus reducing the residual artifacts which can result from the noise reduction process.

5.2 Spectral Subtraction - Boll's Method

In this section, the spectral subtraction algorithm developed by Boll [13,14] is described. Consider an input signal $y(k)$ consisting of a stationary random signal $s(k)$ which has been corrupted by an uncorrelated additive noise source $n(k)$. If the power spectral density of the noise $n(k)$ is known, then it is possible to determine the power spectral density of the signal $s(k)$. That is, given that noise has been added to a signal,

$$y(k) = s(k) + n(k) \quad , \tag{5.1}$$

then the following relation applies,

$$P_y(\omega) = P_s(\omega) + P_n(\omega) \quad , \tag{5.2}$$

where $P_y(\omega)$, $P_s(\omega)$, and $P_n(\omega)$ are the power spectral densities of $y(k)$, $s(k)$ and $n(k)$ respectively. Therefore, an estimate of $P_s(\omega)$ can be obtained by subtracting $P_n(\omega)$ from $P_y(\omega)$. However, because audio signals (including speech) are time-varying processes, the above reasoning must be modified somewhat. Specifically, the observed signal $y(k)$ is windowed into short-time segments. The windowed segment of the observed input signal $y_w(k)$ is obtained by multiplying a segment of $y(k)$ by an appropriate windowing function $w(k)$. Similarly, $s_w(k)$ and $n_w(k)$ are windowed versions of $s(k)$ and $n(k)$ respectively. The reason for using windowed segments of the input signal $y_w(k)$ is that speech can be considered to be locally stationary over periods of about 30 to 40 ms [14,95]. Therefore, by choosing an appropriate window length, $s_w(k)$ can be assumed to be stationary.

Windowing the input signal implies a modified version of equation (5.1), reflecting the fact that the processing must be carried out on a short-time basis,

$$y_w(k) = s_w(k) + n_w(k). \tag{5.3}$$

Taking the Discrete Time Fourier Transform (DTFT) of (5.3) we get,

$$Y_w(e^{j\omega}) = S_w(e^{j\omega}) + N_w(e^{j\omega}) \tag{5.4}$$

where in general for any arbitrary signal $f(k)$ of length L,

$$F_w(e^{j\omega}) = \sum_{k=0}^{L-1} f_w(k) e^{-j\omega k} \tag{5.5a}$$

$$= \sum_{k=0}^{L-1} f(k) w(k) e^{-j\omega k}. \tag{5.5b}$$

Similarly, for any $F_w(e^{j\omega})$, the inverse Discrete Time Fourier Transform is found by

$$f_w(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_w(e^{j\omega}) e^{j\omega k} d\omega. \tag{5.6}$$

Equations (5.3), (5.4), (5.5), and (5.6) therefore provide the following transform pairs:

$$y_w(k) \leftrightarrow Y_w(e^{j\omega}), \tag{5.7a}$$

$$s_w(k) \leftrightarrow S_w(e^{j\omega}), \tag{5.7b}$$

$$n_w(k) \leftrightarrow N_w(e^{j\omega}). \tag{5.7c}$$

In the spectral subtraction technique proposed by Boll it is assumed that the spectral magnitude of the noisy speech can be successfully approximated by the sum of the speech and noise spectral magnitudes. That is, Boll made the approximation,

$$\left|Y_w(e^{j\omega})\right| \approx \left|S_w(e^{j\omega})\right| + \left|N_w(e^{j\omega})\right| \tag{5.8}$$

from which the spectral magnitude of the speech signal can be estimated as

$$\left|\hat{S}_w(e^{j\omega})\right| = \left|Y_w(e^{j\omega})\right| - \left|N_w(e^{j\omega})\right|, \tag{5.9}$$

where $\left|\hat{S}_w(e^{j\omega})\right|$ is the estimated spectral magnitude of the clean speech signal.

Since it is assumed that only the degraded signal is available, the magnitude of the noise spectrum $\left|N_w(e^{j\omega})\right|$ is not readily available. Therefore, $\left|N_w(e^{j\omega})\right|$ is approximated by $E\left[\left|N_w(e^{j\omega})\right|\right]$, where $E[\cdot]$ denotes the expectation operator. Typically, $n_w(k)$ is as-

sumed to be locally stationary in the sense that the spectral magnitude of the noise just prior to speech activity is the same as during speech activity. Furthermore, $n_w(k)$ is assumed to be ergodic, and so that in practice $E\left[\left|N_w(e^{j\omega})\right|\right]$ is obtained by averaging measurements of $\left|Y_w(e^{j\omega})\right|$ during periods with no speech activity, where only the noise is present. As such, the estimated spectral magnitude of the clean speech signal is calculated by

$$\left|\hat{S}_w(e^{j\omega})\right| = \left|Y_w(e^{j\omega})\right| - E\left[\left|N_w(e^{j\omega})\right|\right], \tag{5.10}$$

where $\left|Y_w(e^{j\omega})\right|$ is obtained directly from the observed data.

Equation (5.10) presents an interesting problem in that it can produce a negative estimate for $\left|\hat{S}_w(e^{j\omega})\right|$. Boll suggested two ways of dealing with this problem. One approach is to set all negative values of $\left|\hat{S}_w(e^{j\omega})\right|$ to zero. This is equivalent to half-wave rectification. The second approach is to make all negative magnitudes positive, which is equivalent to full-wave rectification. Therefore, (5.10) can be modified to include half-wave rectification

$$\left|\hat{S}_w(e^{j\omega})\right| = \frac{\left|Y_w(e^{j\omega})\right| - E\left[\left|N_w(e^{j\omega})\right|\right] + \left|\left|Y_w(e^{j\omega})\right| - E\left[\left|N_w(e^{j\omega})\right|\right]\right|}{2}, \tag{5.11}$$

or full-wave rectification,

$$\left|\hat{S}_w(e^{j\omega})\right| = \left|\left|Y_w(e^{j\omega})\right| - E\left[\left|N_w(e^{j\omega})\right|\right]\right|. \tag{5.12}$$

To obtain the noise-reduced signal $\hat{s}_w(k)$, the magnitude $\left|\hat{S}_w(e^{j\omega})\right|$ must be combined with an estimate of the phase of the signal $\Phi_{\hat{S}_w}(e^{j\omega})$ to form $\hat{S}_w(e^{j\omega})$. However, since $\Phi_{S_w}(e^{j\omega})$ is not available, it is replaced by $\Phi_{Y_w}(e^{j\omega})$, the phase of the degraded signal. This approximation is justified because it is well known that listeners tend to be quite insensitive to phase errors in speech over the short term [13,18,19]. The results of subjective tests indicate that listeners do not detect random phase errors of less than about $\pi/4$ over short time intervals. Therefore, the approximation to $s_w(k)$ can be constructed by combining the magnitude and phase estimates and performing an inverse DTFT in the following manner,

$$\Phi_{\hat{S}_w}(e^{j\omega}) \cong \Phi_{Y_w}(e^{j\omega}) \tag{5.13}$$

$$\hat{S}_w(e^{j\omega}) = |\hat{S}_w(e^{j\omega})| e^{\Phi_{Y_w}(e^{j\omega})} \tag{5.14}$$

$$\hat{s}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{S}_w(e^{j\omega}) e^{j\omega k} d\omega. \tag{5.15}$$

The above derivation of the spectral subtraction process can also be viewed diagramatically as shown in Figure 5.1. The observed signal $y(k)$ is windowed into short time segments prior to being transformed via a Discrete Fourier Transform (DFT). In practice this is done using a Fast Fourier Transform (FFT) algorithm. The resulting phase is stored for later use while the spectral magnitude is processed. During periods where there is no speech activity an estimate of the spectral magnitude of the noise is derived. This estimate of the magnitude of the noise is subtracted from the magnitude of the noisy input signal. Any negative values are rectified and combined with the stored phase prior to performing an inverse DFT. The output of the inverse-DFT is an estimate of the clean speech signal.

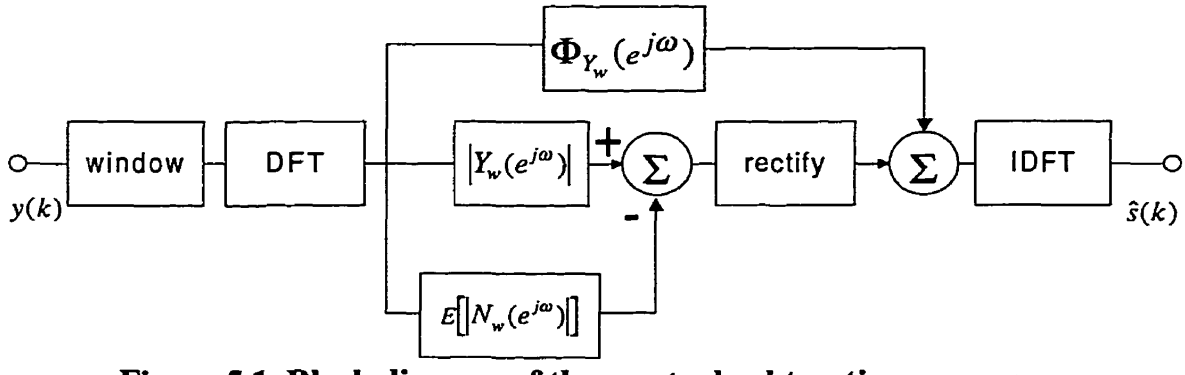


Figure 5.1 Block diagram of the spectral subtraction process.

As stated earlier, the spectral subtraction technique assumes that the noise is locally stationary. If the spectral magnitude of the noise changes to some new locally stationary state, then it is assumed that there is enough time (about 300 ms without speech activity) to determine a new estimate of the noise spectral magnitude. Therefore, for a slowly varying non-stationary background noise source, the algorithm requires some form of speech activity detector in order to know when to update the estimate of the noise spectral magnitude. The estimate of the noise spectral magnitude is obtained by locally averaging the observed signal during periods of non-speech activity. This averaging reduces the variance in the estimate of the noise floor.

Another key assumption is that speech is stationary over short periods of time. Boll chose a window length of 32 ms which is approximately twice the maximum possible

pitch period of the speech. The length of the window must be chosen carefully. A shorter window length will guarantee stationarity of the speech signal but will result in poorer frequency resolution in the spectrum of the noise within a given window. Conversely, a longer window will provide better frequency resolution in the spectral magnitude of the noise but may cause audible artifacts in the reconstructed speech signal since the speech can no longer be assumed to be stationary within the processing frame.

Boll's implementation used windows which were overlapped by 50%. The overlapping greatly reduces any discontinuities in the reconstructed signal that can occur at the boundaries of the windowed segments. Of course, the windows of the processed time data must be overlapped when reconstructing the output time signal. Boll used a Hanning window, while Lim [16] suggested a Bartlett (triangular) window. Lim and Oppenheim [95] suggest that the type of window is not critical to the performance of the algorithm provided that the sum of the overlapping windows is equal to unity as described by the following expression,

$$\sum_i w_i(k) = 1. \tag{5.16}$$

The above discussion relates to the analysis window. There is, of course, a corresponding windowing operation which occurs in the synthesis process. While equation (5.16) implies the use of a rectangular synthesis window, it will be seen later that other analysis/synthesis window combinations can provide improved performance. Moreover, when incorporating an auditory model into the spectral subtraction process, the choice of window becomes important.

Boll used a DFT size equal to the window size. However, Allen [98] points out that modifying the magnitude spectrum in the Fourier domain is equivalent to convolving the signal with a filtering function. Since convolution generally results in a lengthening of the signal, a form of temporal aliasing can result. To eliminate this aliasing, zero padding of the time signal can be done prior to the DFT. In this way, any modifications to the spectrum magnitude will spill into the zero padding upon doing the inverse DFT. Boll found that for his work (voice communication systems), augmenting the signal with zeros did not result in a significant improvement in audio quality.

To test the performance of his algorithm, Boll conducted a DRT (Diagnostic Rhyme Test) to measure the intelligibility of the reconstructed signal. He also conducted tests to evaluate the subjective quality of the processed waveform. The signal-to-noise ratio of the unprocessed speech was in the range of -5 dB to +5 dB, and the results indicated that

the spectral subtraction algorithm had no positive effect on intelligibility. However, the *perceived* quality of the speech was significantly improved.

Many variations to the spectral subtraction algorithm have been developed. The various algorithms differ primarily in the manner in which they estimate $|\hat{S}_w(e^{j\omega})|$. Many alternate spectral subtraction algorithms can be derived through a generalization of (5.10) as shown below,

$$|\hat{S}_w(e^{j\omega})|^\alpha = |Y_w(e^{j\omega})|^\alpha - \beta \cdot E\left[|N_w(e^{j\omega})|^\alpha\right] \tag{5.17}$$

where $\alpha > 0$ and β is the overestimation parameter. Setting both parameters, α and β equal to unity results in the magnitude subtraction algorithm proposed by Boll. Setting α equal to 2 results in the power subtraction version of the spectral subtraction algorithm [99]. Lim [16] conducted a study in which he investigated the effect of α on speech intelligibility. He performed intelligibility tests with 22 listeners on speech segments having signal-to-noise ratios of ∞, +5, 0, and -5 dB, processed using the spectral subtraction algorithm described by (5.17) with values of 2.0, 1.0, 0.5 and 0.25 for α (β=1). Lim found that within the range from 0.5 to 2.0 the choice of α did not significantly affect the measured intelligibility of the processed speech. For α =0.25 the intelligibility scores decreased substantially. Interestingly however, Lim indicates that for values of α of 1.0 and 0.5, the processed speech was perceived to be "less noisy". The overestimation parameter β allows for the possibility of subtracting more than the expected value of the noise and will be discussed in greater detail in later sections.

5.3 Interpretation of Spectral Subtraction as a Zero-Phase Filter

Lim and Oppenheim [95] showed that the spectral subtraction algorithm can be viewed as a zero-phase filter. That is, the spectral subtraction process can be described by the following expression,

$$\hat{S}_w(e^{j\omega}) = Y_w(e^{j\omega}) \cdot H(e^{j\omega}). \tag{5.18}$$

For example, substituting (5.17) into (5.18) and setting α equal to 2 provides the following expression for $H(e^{j\omega})$,

$$H(e^{j\omega}) = \frac{\hat{S}_w(e^{j\omega})}{Y_w(e^{j\omega})} = \left(\frac{|Y_w(e^{j\omega})|^2 - \beta \cdot E\left[|N_w(e^{j\omega})|^2\right]}{|Y_w(e^{j\omega})|^2} \right)^{1/2} . \tag{5.19}$$

Furthermore, by defining

$$X^2(e^{j\omega}) = \frac{|Y_w(e^{j\omega})|^2}{E\left[|N_w(e^{j\omega})|^2\right]} , \tag{5.20}$$

and substituting this into (5.19), the following expression for $H(e^{j\omega})$ is obtained,

$$H(e^{j\omega}) = \left(\frac{X^2(e^{j\omega}) - \beta}{X^2(e^{j\omega})} \right)^{1/2} . \tag{5.21}$$

The parameter $X^2(e^{j\omega})$ as defined in (5.20) is the signal-plus-noise-to-noise ratio at each frequency ω. The filter described by (5.21) can be described as a zero-phase filter since it uses only magnitudes and thus has no effect on the phase of the signal. Figure 5.2 plots $H(e^{j\omega})$ as a function of $X^2(e^{j\omega})$ for values of β of 0, 1, 2 and 4.

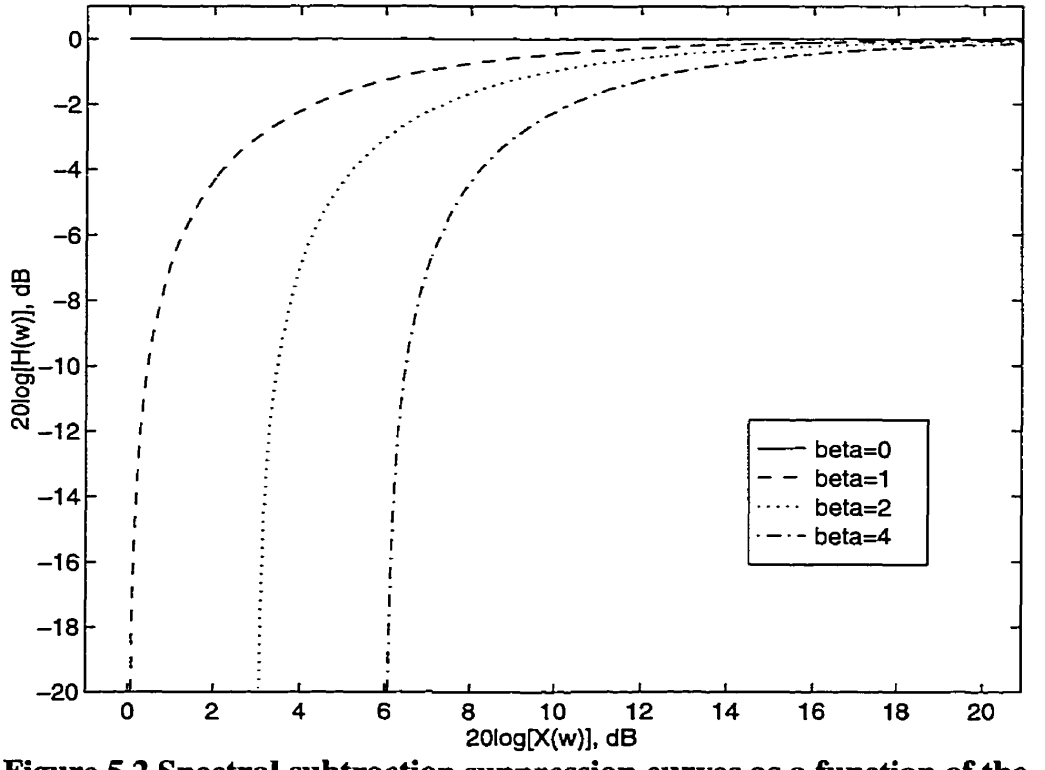


Figure 5.2 Spectral subtraction suppression curves as a function of the overestimation parameter, β (α=2).

The figure reveals that the spectral subtraction process, at a frequency ω, can be described as a family of attenuation (suppression) curves which are dependent upon the value of β. The amount of attenuation applied to the noisy signal, and therefore the amount of noise suppression, varies with the local signal-plus-noise-to-noise ratio of the unprocessed signal. At higher values of $X(e^{j\omega})$, (i.e. ≥ 20 dB) very little processing is required and therefore almost no attenuation is applied to the noisy signal. As the signal-to-noise ratio decreases however, more noise suppression is required and so more attenuation is applied to the signal. It can also be seen from the figure that increasing the overestimation factor β causes more attenuation to be applied to the noisy signal. More precisely, the suppression curves are offset (horizontally) by 3 dB for every doubling of β.

The zero-phase filter interpretation of the spectral subtraction process can be generalized beyond what Lim and Oppenheim described (see equation (5.19)). Specifically, by substituting (5.17) into (5.18) without assuming a specific value for α, the following equation is obtained,

$$H(e^{j\omega}) = \frac{\hat{S}_w(e^{j\omega})}{Y_w(e^{j\omega})} = \left(\frac{|Y_w(e^{j\omega})|^{\alpha} - \beta \cdot E\left[|N_w(e^{j\omega})|^{\alpha} \right]}{|Y_w(e^{j\omega})|^{\alpha}} \right)^{1/\alpha} . \tag{5.22}$$

By now defining

$$X^{\alpha}(e^{j\omega}) = \frac{|Y_w(e^{j\omega})|^{\alpha}}{E\left[|N_w(e^{j\omega})|^{\alpha} \right]} , \tag{5.23}$$

and substituting $X^{\alpha}(e^{j\omega})$ into (5.22), the following expression for $H(e^{j\omega})$ is obtained,

$$H(e^{j\omega}) = \left(\frac{X^{\alpha}(e^{j\omega}) - \beta}{X^{\alpha}(e^{j\omega})} \right)^{1/\alpha} . \tag{5.24}$$

A further generalization can be made by allowing the exponent outside the parenthesis of (5.24) to be independent of the exponent inside the parenthesis. More precisely, a new variable γ is defined and (5.24) is modified to give,

$$H(e^{j\omega}) = \left(\frac{X^{\alpha}(e^{j\omega}) - \beta}{X^{\alpha}(e^{j\omega})} \right)^{1/\gamma} . \tag{5.25}$$

In a study by Paul [100], this additional degree of freedom was found to be helpful in allowing for a balance between the amount of noise suppression and the amount of signal distortion resulting from the processing.

Expressed in the form of equation (5.17), this generalized version of spectral subtraction is given as,

$$\left|\hat{S}_w(e^{j\omega})\right|^{\gamma} = \left|Y_w(e^{j\omega})\right|^{\alpha} - \beta \cdot E\left[\left|N_w(e^{j\omega})\right|^{\alpha}\right]. \tag{5.26}$$

It should be noted that, for values of $\alpha \neq 2$, $X^{\alpha}(e^{j\omega})$ as defined in (5.23) is no longer equal to the signal-plus-noise-to-noise ratio. In order to plot $H(e^{j\omega})$ as a function of the signal-plus-noise-to-noise ratio, the parameter $X^{\alpha}(e^{j\omega})$ must be described in terms of $X^2(e^{j\omega})$ as shown below,

$$X^{\alpha}(e^{j\omega}) = \left(X^2(e^{j\omega})\right)^{\alpha/2}. \tag{5.27}$$

The parameter $X^2(e^{j\omega})$ represents the signal-plus-noise-to-noise ratio of the unprocessed signal. A parameter which is perhaps more convenient for understanding the properties of the zero-phase spectral subtraction filter is the traditional signal-to-noise ratio denoted here as $R(e^{j\omega})$, and related to $X^2(e^{j\omega})$ through the following,

$$X^2(e^{j\omega}) = R(e^{j\omega}) + 1. \tag{5.28}$$

Given (5.25) and (5.28), it is possible to examine the suppression curves for the more generalized spectral subtraction algorithm. Examples of various suppression curves resulting from tradeoffs of the parameters α, β and γ are given in Figures 5.3 to 5.6.

Figure 5.3 shows a family of suppression curves which were obtained by varying α and γ together, while keeping β equal to unity. Note that the horizontal axis represents the signal-to-noise ratio $R(e^{j\omega})$. It can be seen that more attenuation is applied to the noisy signal for lower values of α (and γ) and hence more noise reduction is obtained. The curve with $\alpha = \gamma = 1$ represents the magnitude subtraction algorithm proposed by Boll, while the curve with $\alpha = \gamma = 2$ is the suppression curve for the power subtraction algorithm. Clearly the magnitude subtraction algorithm represents a more aggressive noise reduction algorithm than the power subtraction approach.

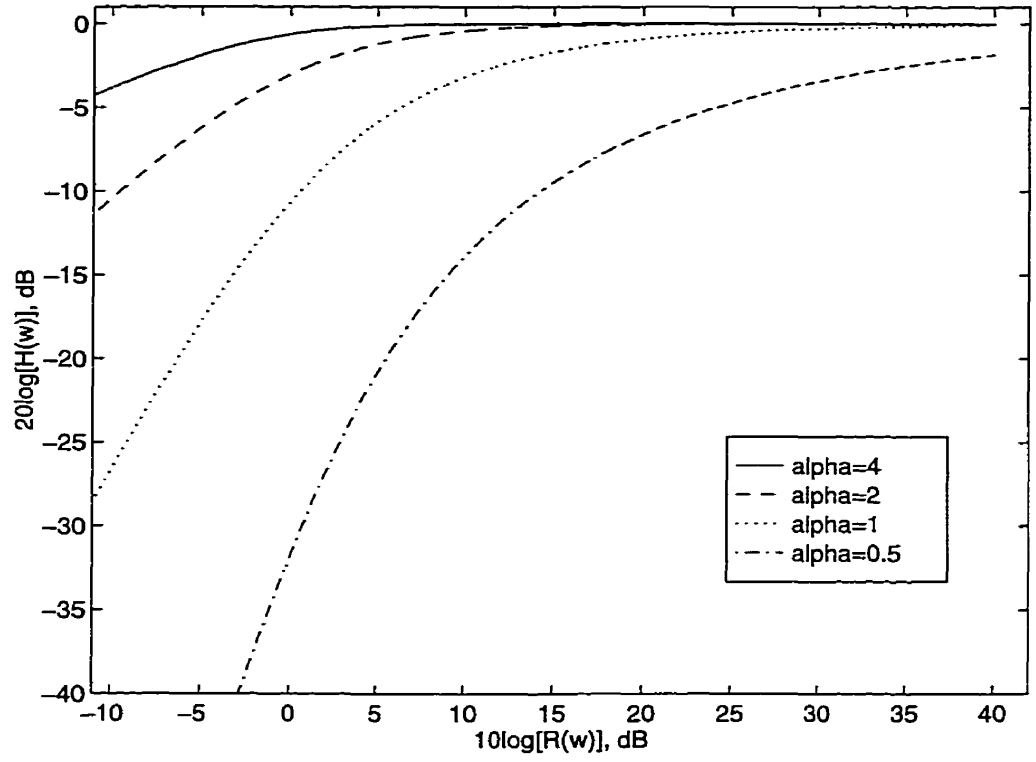


Figure 5.3 Suppression curves for four values of α, with $\gamma=\alpha$ and $\beta=1$.

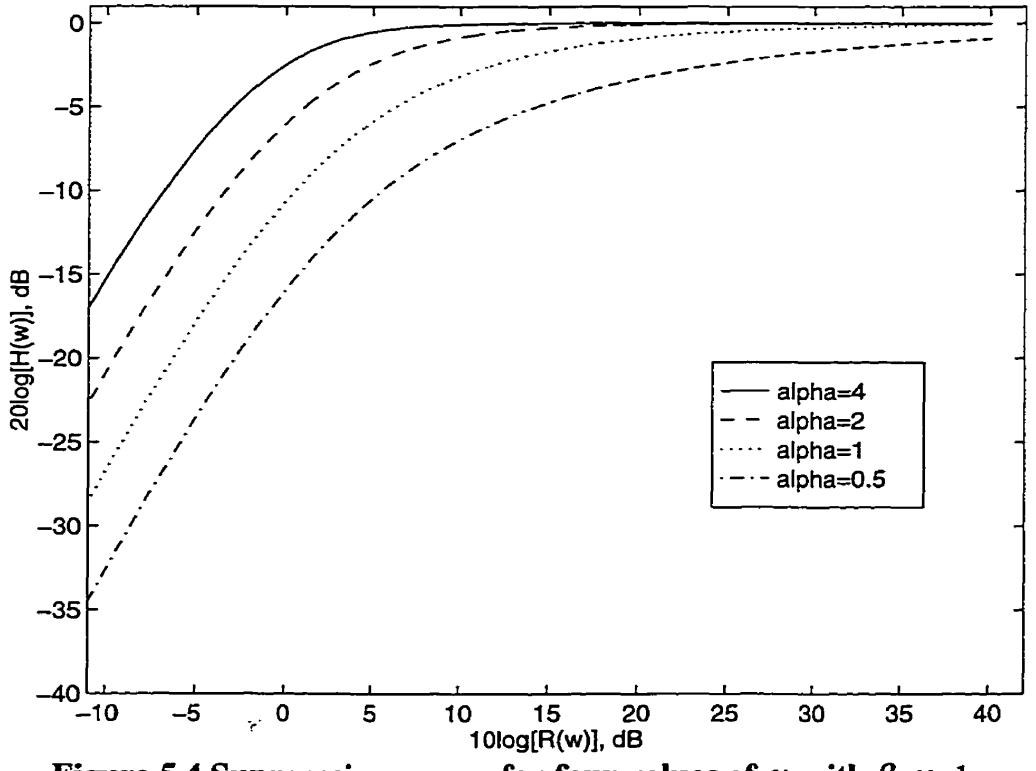


Figure 5.4 Suppression curves for four values of α, with $\beta=\gamma=1$.

The curves of Figure 5.4 show the effect of changing α while keeping $\gamma = 1$. Again, β is held equal to unity for all curves. For signal-to-noise ratios below 0 dB, the suppression curves become linear and are parallel to each other. Furthermore, these linear portions of the suppression curves are offset vertically by 6 dB for every doubling (or halving) of α.

The curve corresponding to $\alpha = 2$ (with $\gamma = 1$) is the Wiener filter version of the spectral subtraction algorithm. That is, it approximates a Wiener filter and stems from an attempt to minimize the mean-squared error of the best time domain fit to the underlying speech signal. Similarly, the power subtraction algorithm described earlier corresponds to the best estimate of the spectrum of the speech [15].

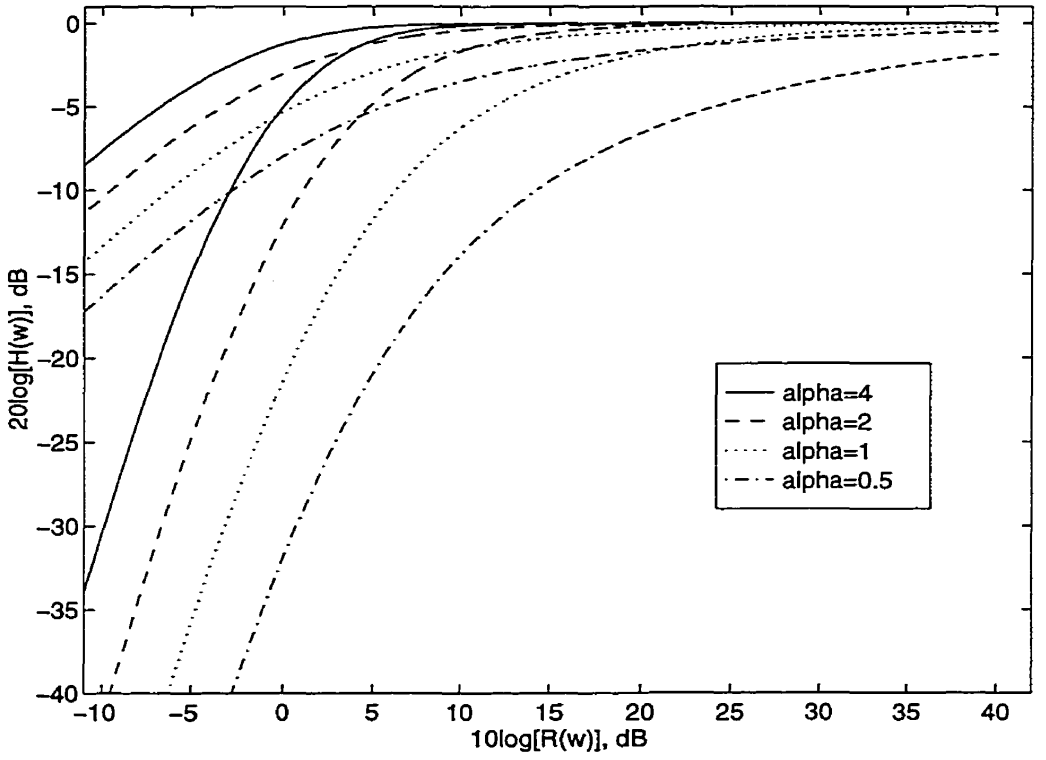


Figure 5.5 Suppression curves for four values of α and two values of γ. Upper curves are for γ=2. Lower curves are for γ=0.5.

The curves in Figure 5.5 represent two groups of suppression curves which are obtained for two values of γ. The upper group of curves correspond to a value of $\gamma = 2$ while the lower set of curves are for $\gamma = 0.5$. Within each group the value of α is varied in four steps from 0.5 to 4. Lowering the value of γ provides a sharper knee in the attenuation curve and results in a steeper slope at lower signal-to-noise ratios. Note however, that within the two groups the linear portions of the curves remain parallel. That is,

the curves within a group have the same slope for signal-to-noise ratios below approximately 0 dB.

Figure 5.6 demonstrates the effect of varying γ. For the suppression curves shown in this figure, α and β are held constant at unity. Examination of the linear portions of the curves (i.e. $R(e^{j\omega}) \leq 0$ dB) shows that γcontrols the slope of the curves. Specifically, a halving of γresults in a doubling of the slope.

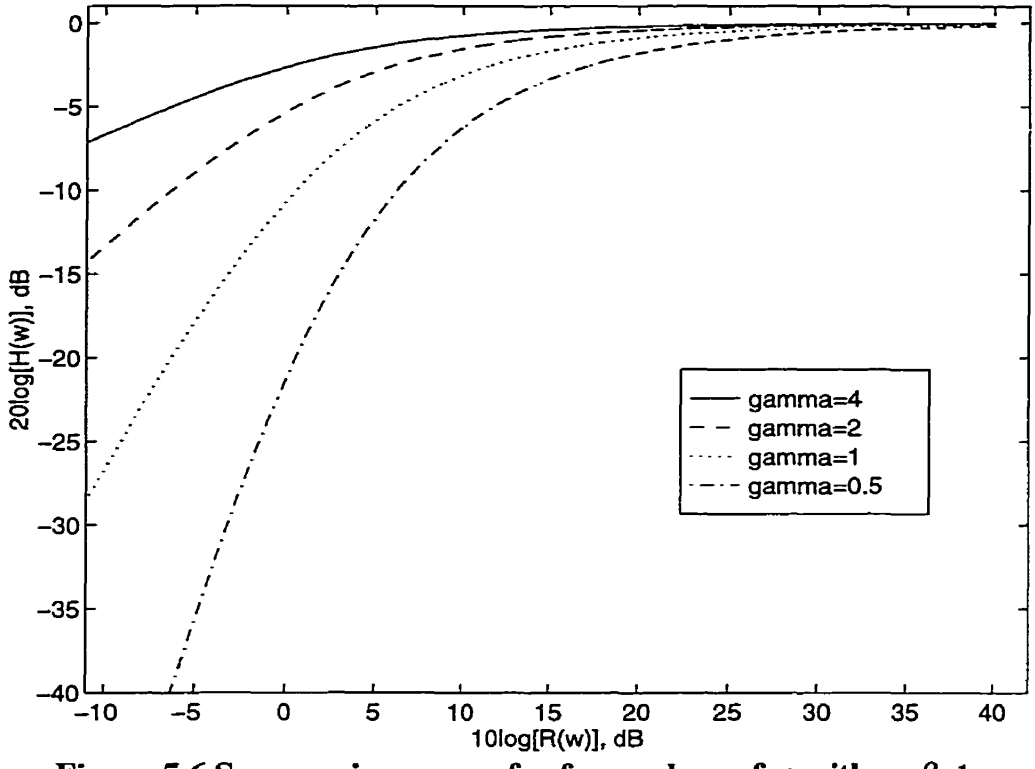


Figure 5.6 Suppression curves for four values of γ, with α=β=1.

Given the results of Figures 5.4 and 5.6, the combined effect of changing both α and γ can be understood. Referring back to Figure 5.3, consider the linear portions of the curves for $\alpha = \gamma = 1$ and $\alpha = \gamma = 2$. Examination of these curves at $R(e^{j\omega}) = -10$ dB, shows that the attenuation of the $\alpha = \gamma = 1$ curve can be derived by doubling the slope of the $\alpha = \gamma = 2$ curve (due to the halving of γ) and subtracting 6 dB (due to the halving of α).

The results plotted in Figures 5.2 to 5.6 demonstrate that considerable flexibility in controlling the suppression curves is possible through changes in the parameters α, β and $\gamma$of (5.25). Suitable values for these parameters depend on the signal-to-noise ratio of the unprocessed signal, the required amount of noise suppression and the acceptable level of

distortion to the underlying signal. It has also been shown in Figures 5.2 to 5.6 that the slope of the linear portions of these curves is directly controlled by α, β and γ. This is an important consideration since it has been shown by Paul [100] that the points at which the various spectral subtraction algorithms create unpleasant artifacts tend to have the same slope on the noise suppression curves. That is, when the slope of the suppression curves exceed a critical value, perceptually disturbing artifacts appear.

5.4 Limitations of the Spectral Magnitude Estimation Methods

In the previous sections the fundamental principles of signal enhancement based on estimating the spectral magnitude were described. In this section, the performance of this technique is discussed and its limitations are identified.

It has already been noted that spectral subtraction does not tend to improve speech intelligibility, but does provide an improvement in the perceived quality of the signal. However, its performance is limited by the audible artifacts created by the process which can be more disturbing to some listeners than the original noise. These artifacts become increasingly disturbing as the signal-to-noise ratio of the corrupted input signal decreases. In order for the spectral subtraction technique to be useful for reducing camera noise, steps must be taken to reduce the audibility of these artifacts while providing a sufficient amount of noise suppression.

The various types of artifacts created by the spectral subtraction process include:

- musical noise,
- incomplete or variable cancellation of the noise (modulation of the noise floor),
- timbral effects and/or loss of frequency components of the signal,
- missing sounds - loss of low level signal (speech) components,
- phase distortions,
- time aliasing,
- pre-echoes and post-echoes (temporal smearing).

All of these artifacts are due to errors in two of the underlying assumptions upon which spectral subtraction is based. Specifically, the artifacts can be shown to be due to errors in the assumption that the spectral magnitude of the noise (within a given processing frame) is equal to the expected value of the noise,

$$\left|N_w(e^{j\omega})\right| \approx E\left[\left|N_w(e^{j\omega})\right|\right],$$

as well as errors in approximating the phase of the underlying signal by the phase of the noisy signal,

$$\Phi_{\hat{S}_w}(e^{j\omega}) \approx \Phi_{Y_w}(e^{j\omega}).$$

5.4.1 Musical Noise

Musical noise is possibly the single most limiting factor when using the spectral subtraction process for removing camera noise. Musical noise consists of short spurious bursts of isolated frequency components which appear seemingly at random across the spectrum of the processed signal. As noted earlier, because of the time-varying nature of speech, spectral subtraction must be done on a frame-by-frame basis. Due to the random frame-to-frame fluctuations in the magnitude spectrum of the noise, the level of a given spectral component will sometimes be greater than the estimated value of the noise. Therefore, the spectral subtraction process will not entirely cancel these components and short tone bursts will exist for the duration of the frame. Because of its musicality (tonality), this residual noise is often very disturbing to the listener and for the purpose of removing camera noise, musical noise must be strictly avoided. An example of musical noise is provided in Figure 5.7 which shows the spectrogram of a signal with no speech activity processed by spectral subtraction (with $\alpha = \beta = \gamma = 1$). The musical noise is seen as the darker areas (rectangles) in the figure which represent short bursts of residual narrowband noise.

It should be noted that musical noise occurs more when half-wave rectification is used in the spectral subtraction process. This is because with half-wave rectification, variations in the level of a given spectral component can result in that component being randomly switched on and off on a frame-by-frame basis. Conversely, with full-wave rectification, negative spectral components become positive and no sudden switching occurs. However, full-wave rectification limits the amount of noise suppression that is possible, and can in fact increase the noise level in some instances [13]. Therefore, developers of spectral subtraction algorithms have used half-wave rectification exclusively.

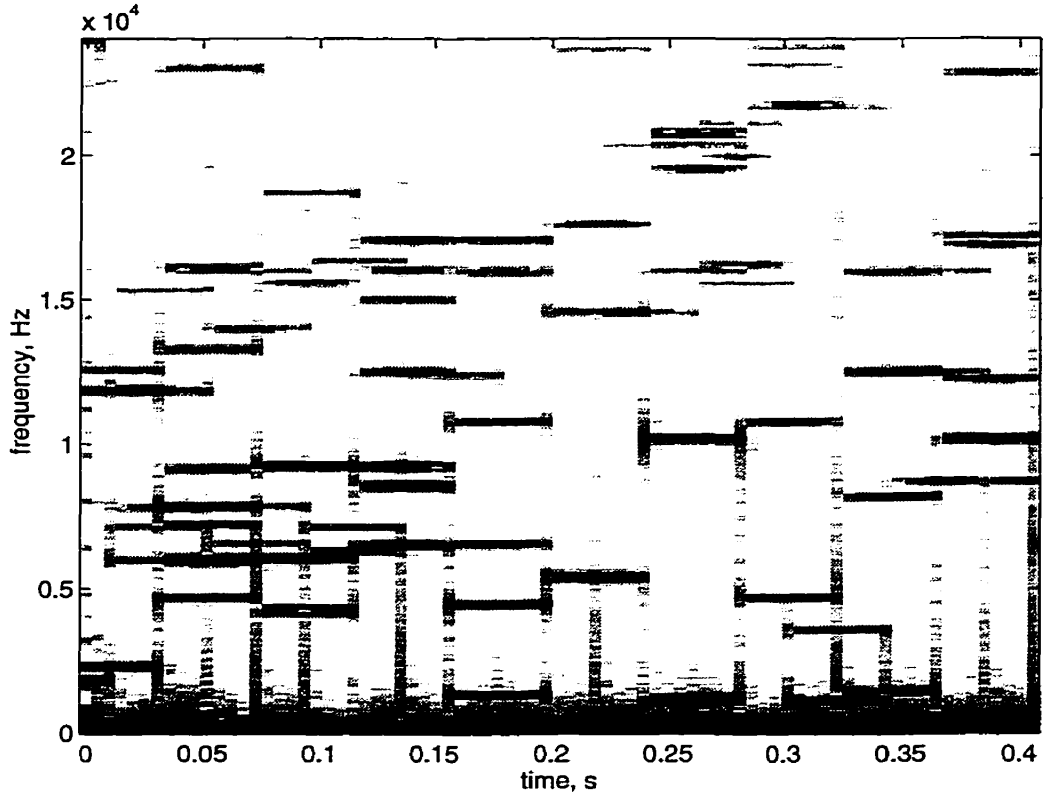


Figure 5.7 Spectrogram of processed signal showing musical noise.

5.4.2 Ephraim and Malah's Spectral Subtraction Algorithm

Ephraim and Malah [22] proposed a version of spectral subtraction which uses a minimum mean-square error estimate of the magnitude spectrum of the noise. The main point of interest regarding this algorithm is that it produces *colorless* residual noise and does not create musical noise as a result of its processing. This is achieved by using the concept of an *a priori* signal-to-noise estimate SNR_{prio} which is defined as,

$$SNR_{prio}(p,e^{j\omega})=(1-\varphi)P[SNR_{post}(p,e^{j\omega})]+\varphi\frac{|\hat{S}(p-1,e^{j\omega})|^2}{E[|N(e^{j\omega})|^2]}\ , \tag{5.29}$$

with

$$SNR_{post}(p,e^{j\omega})=\frac{|Y(p,e^{j\omega})|^2}{E[|N(e^{j\omega})|^2]}\ , \tag{5.30}$$

where $P[\cdot]$ denotes half-wave rectification. $SNR_{post}(p,e^{j\omega})$ is the *a posteriori* estimate of the signal-to-noise ratio in the pth time interval, while $|\hat{S}(p-1,e^{j\omega})|^2$ is the estimate of the desired signal in the p-1 time interval. The parameter φ determines the amount of smoothing applied in the estimate of SNR_{prio} and is typically set to about 0.98.

In an effort to determine the mechanism within the algorithm which causes the residual noise to be colorless, Cappé [21] conducted a comprehensive study of the Ephraim and Malah noise suppressor and concluded that the nonlinear smoothing procedure limits the variation in the attenuation applied to the noisy signal over successive frames. By limiting the frame-to-frame variation in the attenuation, the on-and-off switching of spectral components responsible for musical noise is correspondingly limited. Cappé's findings tend to support the observations of Paul [100] who noted that the onset of musical noise occurs when the slope of the noise suppression curves exceeds a critical value.

As the slope of the suppression curve increases, the variation in the level of a given spectral component increases relative to the average level. Paul speculated that musical noise becomes audible when a critical value of this ratio is exceeded. Stated more simply, a steeper slope on the suppression curve tends to emphasize the on-and-off switching of low level spectral components. When the slope exceeds a certain value, the switching increases to the point that it becomes audible. Recently, Scalart and Vieira Filho [101] showed that the smoothed noise estimate technique can be used to significantly reduce musical noise in the various spectral subtraction algorithms (magnitude, power, Wiener, etc.) described earlier.

The Ephraim and Malah noise suppressor is not directly applicable to the problem of camera noise since it provides only a reduction of the interfering noise and does not give complete cancellation of the noise. In some applications, such as the restoration of gramophone recordings, complete noise cancellation is not required and so the Ephraim and Malah noise suppresser can be applied [30]. For the problem of camera noise however, other schemes must be investigated for reducing the musical noise.

5.4.3 Signal Subspace Approach

More recently, Ephraim and Van Trees [102,103] proposed a signal subspace approach to speech enhancement. They decomposed the vector space of the noisy signal into two subspaces; a signal plus noise subspace and a noise subspace. The noise reduction process then consists of nulling the noise subspace and estimating the signal from the remaining subspace. The main difference between the signal subspace approach and the spectral subtraction method is in the transform used to decompose the noisy signal. The Karhunen-Loève transform (KLT) was used for the decomposition in the signal subspace approach.

As in the Ephraim and Malah method described above, the main advantage claimed for the signal subspace approach is that it does not create musical noise. To evaluate the performance of the signal subspace approach, Ephraim and Van Trees conducted two subjective tests. In the first test they found that 84% of the listeners preferred the speech enhanced by the signal subspace method over the unprocessed noisy speech. The remaining listeners felt that the distortion of the speech signal due to the processing was more disturbing than the noise. In the second test, listeners compared the amount of distortion to the speech signal resulting from the signal subspace approach and the spectral subtraction approach. The results indicate that the signal subspace approach causes more audible distortion of the speech signal than the spectral subtraction method.

Therefore, the signal subspace approach may not be appropriate for the camera noise application since distortion to the underlying speech signal must be strictly avoided. Moreover, while the main benefit of the signal subspace approach appears to be that it does not produce musical noise, we shall see later in this chapter that spectral subtraction methods based on auditory perception can also reduce musical noise while minimizing distortions to the speech signal.

5.4.4 Wavelet Based Noise Reduction

A noise reduction method which shares many similarities with the signal subspace method described above is the method based on wavelets [104,105].

In this approach, the noisy signal is expanded in an appropriate orthonormal basis. The choice of basis is made using some form of cost function such as the Shannon entropy. The coefficients of the expansion are ordered in terms of magnitude. Those coefficients which fall above some pre-determined threshold are assumed to be due the coherent (desired) portion of the input signal. The residual terms consist of the noisy part of the input signal and are treated as a new signal which is in turn expanded and divided into its coherent and noisy components. This iteration process continues and the coherent portions from each expansion are combined to produce an estimate of the clean signal.

Berger *et al.* [104] used this approach to restore old musical recordings of piano and vocal arrangements. They report that while the wavelet based denoising algorithm was useful for removing noise from musical signals, it created several undesirable artifacts in the restored signal. Specifically, they report disturbing signal-dependent fluctuations in the level of the residual background noise (i.e., noise spurts). Among other artifacts,

Berger *et al.* also indicate that the method creates annoying random clicks and whistles (tones) in the restored signal.

It appears that while wavelet based noise reduction may offer an alternative method for noise reduction, it has several shortcomings which limit its performance. Therefore, in this thesis, we shall restrict our discussion to spectral subtraction based algorithms. Later in this chapter, the similarities between the concept of decomposing the spectral subtraction process into subbands and sub-frames, and wavelet decomposition will be highlighted. An advantage of the subband/sub-frame approach is that it can be easily extended to include a model of the human auditory system which in turn can significantly reduce the audibility of unwanted artifacts and distortion to the speech signal.

5.4.5 Overestimation with Minimum Spectral Floor

Many schemes have been developed to try to avoid or eliminate musical noise. The earliest attempts used the overestimation parameter defined earlier as β in (5.17) [17]. By setting $\beta > 1$, the level of the noise is overestimated and thus the amount of noise suppression is increased. This in turn, reduces the amount of musical noise in the processed signal by limiting the number of spectral components that go uncanceled. By progressively increasing β, one can reduce the musical noise to an arbitrary level. In practice, values in the range from 1.5 to 2.5 are used for Boll's version of spectral subtraction.

Rather than multiplying the expected value of the noise magnitude by an overestimation parameter, Preuss [27] proposed using the maximum value of the spectral magnitude of the noise measured during periods with no speech component. This is effectively a form of overestimation. More recently, Lockwood and Boudy [24] suggested that the overestimation parameter should be frequency dependent.

The problem with any approach for reducing musical noise based on some form of overestimation is that one inevitably cancels a portion of the desired speech signal. This becomes increasingly true at lower signal-to-noise ratios where distortions to the underlying speech signal become readily audible. This distortion is not acceptable when removing camera noise. At higher signal-to-noise ratios, a significant amount of overestimation is possible before any distortion to the speech signal becomes audible and therefore, under these conditions, the musical noise can be effectively eliminated. At lower signal-to-noise ratios a modest overestimation is acceptable but must be combined with other techniques in order to remove the musical noise.

As a means of providing a tradeoff between the amount of attainable noise suppression, the audibility of musical noise, and the level of distortion to the underlying speech signal, Berouti *et al.* [17] proposed the use of a *minimum spectral floor.* In this scheme, any spectral component falling below some threshold is set to a value referred to as the spectral floor. Mathematically, the process can be described by modifying (5.26) as follows;

$$D_w(e^{j\omega}) = \left|Y_w(e^{j\omega})\right|^{\alpha} - \beta \cdot \left|N_w(e^{j\omega})\right|^{\alpha} , \tag{5.31}$$

$$\left|\hat{S}_w(e^{j\omega})\right|^{\gamma} = \begin{cases} D_w(e^{j\omega}), & D_w(e^{j\omega}) > \delta \cdot \left|N_w(e^{j\omega})\right|^{\alpha} \\ \delta \cdot \left|N_w(e^{j\omega})\right|^{\alpha}, & \textit{otherwise} \end{cases} , \tag{5.32}$$

where $\left|\hat{S}_w(e^{j\omega})\right|$ is the estimate of the spectral magnitude of the desired speech signal.

The parameter $\delta \cdot \left|N_w(e^{j\omega})\right|^{\alpha}$ corresponds to the spectral floor. The parameter δ limits the amount of noise suppression by keeping the residual noise above some minimum level. This in turn limits the variance of the residual spectral components, thus limiting the on-and-off switching of these components which causes the musical noise. Smaller values of δ provide a greater degree of noise suppression but result in more audible musical noise. Experiments have shown that the optimal value for δ is dependent on α, β, and γ[17, 100].

As an extension to the minimum spectral floor method for the camera noise case, we propose the following. Due to the characteristics of the camera noise, specifically the periodic component, there is some structure to the phase of the camera noise. Therefore, when implementing a minimum spectral floor, the structure in the phase is maintained and the residual noise is clearly audible as camera noise. To overcome this problem, the phases of the components which fall below the minimum spectral floor are set to some random value. The values for the phase are chosen from a uniformly distributed random process. As can be heard on the demonstration CD, this results in a more benign (hiss-like) noise which may be more acceptable in the camera noise application.

5.4.6 Survival Algorithm

The approach of the methods presented so far for reducing musical noise has been to integrate some scheme directly into the spectral subtraction process. A somewhat different approach was proposed by Vaseghi and Frayling-Cork [20] whereby the output of the

spectral subtraction algorithm is processed in a separate algorithm designed specifically to reduce musical noise.

The overestimation schemes discussed in the previous section considered only the magnitude of the spectral peaks which cause musical noise. The Vaseghi and Frayling-Cork algorithm takes into account the duration of the spectral peaks as well as their level. They found that a large proportion of the spectral peaks of the noise which result in musical noise had a duration of less than 15 ms, whereas spectral components due to the speech signal tend to have a longer lifetime. Therefore, they reasoned that a means of reducing musical noise could be devised based on the lifetime of a given spectral peak.

Vaseghi and Frayling-Cork proposed a *survival algorithm* which passes a "window" over the successive frames of the spectra resulting from the output of the spectral subtraction process for each of the frequency bins. The size of the window was chosen to accommodate 5 frames from the spectral subtraction process. The extreme end frames in the window are tested to determine if they are simultaneously zero. Having both end frames equal to zero indicates that the spectral components contained within the window are of short duration. If the two end frames of a given frequency bin are not simultaneously zero, then this suggests that a desired signal component is contained within the window and no processing is done on these frames. If however, they are equal to zero, then a second test is performed to see if any of the frames within the window exceed a threshold level which would indicate the presence of a signal. If any of the frames exceeds this threshold then no processing is done to the frames. If none of the frames exceed the threshold, then all of the frames within the window are set to zero. The window is then shifted to the next group of 5 frames and the process is repeated.

While this method is somewhat heuristic in its approach, Vaseghi and Frayling-Cork reported that their survival algorithm removes the majority of musical noise and results in considerably lower residual noise energy and a substantial improvement in perceived quality.

5.4.7 Modified Survival Algorithm

The survival algorithm described above was implemented and evaluated as a means of reducing musical noise when suppressing camera noise. The size of the window used in the survival algorithm was varied to include from 5 to 10 frames. The threshold used to determine the presence of a signal component was frequency dependent and based on the

estimate of the noise. Specifically, the threshold T was a function of the noise estimate as described by,

$$T = \varphi \cdot E\left[\left|N_w\left(e^{j\omega}\right)\right|\right]. \tag{5.33}$$

Using φ, the threshold could be varied and it was found that the value of T giving the best performance varied with the signal-to-noise ratio of the input signal. As indicated by Vaseghi and Frayling-Cork the algorithm provided a substantial reduction of the musical noise. However, it was found that with some modifications, the algorithm could be made to perform better.

The Vaseghi and Frayling-Cork algorithm makes a binary decision as to whether or not the spectral components within the analysis window are due to musical noise. This decision is based largely on the presence of components at the boundaries of the analysis window. As such, the basic survival algorithm will not eliminate low level musical noise which happens to fall on the boundaries of the analysis window. Similarly, the algorithm does not address musical noise which falls within the same analysis window as a signal component which exceeds the threshold. A modified survival algorithm was therefore devised which attempts to address these issues. The workings of the modified survival algorithm are seen in Figure 5.8.

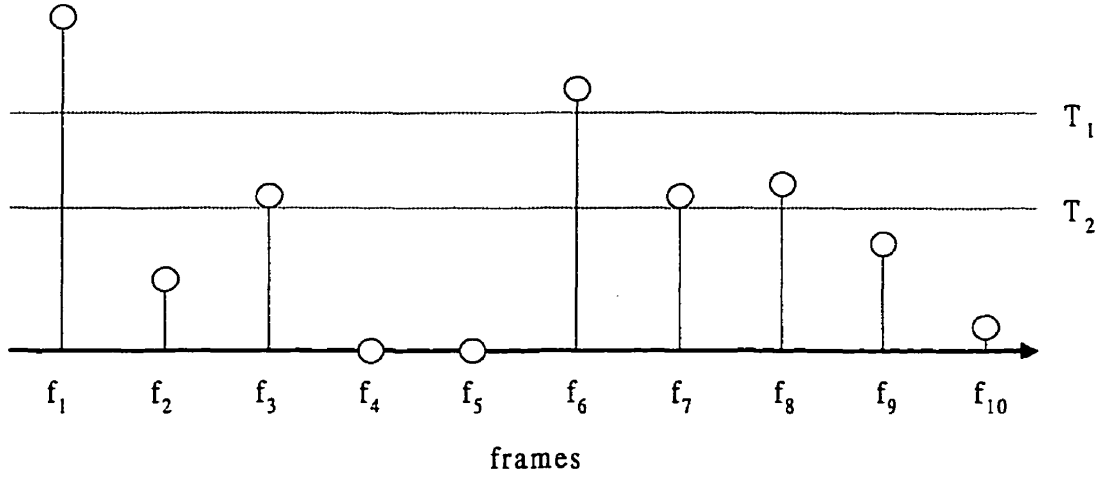


Figure 5.8 Modified survival algorithm for removing musical noise.

As before, an analysis window containing N frames from the output of the spectral subtraction process is used in the proposed modified survival algorithm. The new algorithm uses two thresholds and a soft-decision rule for attenuating low level components. A gain factor, g is applied to each frame $f_i(e^{j\omega})$ as determined by the following rule,

$$f_i(e^{j\omega}) = \begin{cases} f_i(e^{j\omega}) & , \; f_i(e^{j\omega}) \geq T_1(e^{j\omega}) \\ g \cdot f_i(e^{j\omega}) & , \; f_i(e^{j\omega}) < T_1(e^{j\omega}) \end{cases} \tag{5.34}$$

where

$$g = \left[\frac{1}{N} \sum_{i}^{N} \text{sgn} \left\{ \frac{\left| f_i(e^{j\omega}) - T_2(e^{j\omega}) \right| + \left(f_i(e^{j\omega}) - T_2(e^{j\omega}) \right)}{2} \right\} \right]^p . \tag{5.35}$$

The rule works in the following manner. Any frame $f_i(e^{j\omega})$ within the analysis window who's level exceeds the threshold, T_1 is left unaltered. T_1 is chosen such that the probability of a noise spectral peak exceeding this threshold is very low. That is, frames exceeding T_1 are with a high probability due to the signal component.

Those frames which fall below T_1 are multiplied by the gain factor (actually attenuation) g. The gain factor g is determined for a given analysis frame by the percentage of frames which exceed the second threshold, T_2. That is, the fewer the number of frames which exceed T_2, the greater the attenuation applied to these frames. The exponent p is used to further control the attenuation applied to these frames. Larger values of p result in more aggressive attenuation of these low level frames. Both thresholds T_1 and T_2 were made to be functions of the noise estimate in a manner similar to equation (5.31). This modified survival algorithm provides more degrees of freedom.

It can be seen that, unlike the Vaseghi and Frayling-Cork algorithm, all frames in the modified version influence the decision as to whether or not the content of the analysis window is musical noise. As a result, a noise spectral peak located in the same analysis window as a wanted signal component will now be attenuated. Also, short bursts of musical noise will now also be attenuated. The number of frames in the analysis window of the Vaseghi and Frayling-Cork algorithm was determined by the expected duration of noise spectral peaks. The new algorithm allows for flexibility in choosing the size of the analysis window.

To better understand the modified survival algorithm, consider the example depicted in Figure 5.8. In this example, frames f_1 and f_6 would pass unprocessed. All other frames would be attenuated by a factor g. Since half of the frames fall below T_2, the gain factor g would be equal to some value $g = 0.5^p$ determined by the parameter p. In the Vaseghi and Frayling-Cork algorithm, none of the frames in this example would be processed (attenuated).

5.4.8 Incomplete noise cancellation

The basic spectral subtraction algorithm described in Section 5.2 does not provide complete cancellation of the noise. That is, a residual noise will remain after processing. The incomplete cancellation of the interfering noise is due to the variations in the spectral magnitude of the noise. Specifically, the spectral magnitude of the noise varies about some mean value from frame to frame. This is in evidence in Figure 5.9 which shows the average value (dotted curve) of the noise for each spectral component as well as the maximum value (solid curve) over 20 frames. It can be seen that in this example, the maximum value of the noise can exceed the average value by as much as 5 dB. Other examples may demonstrate a larger variation.

In Boll's basic algorithm the expected value of the noise is used in the subtraction process, and so, for those frames where the noise exceeds the expected value, a portion of the noise will not be canceled.

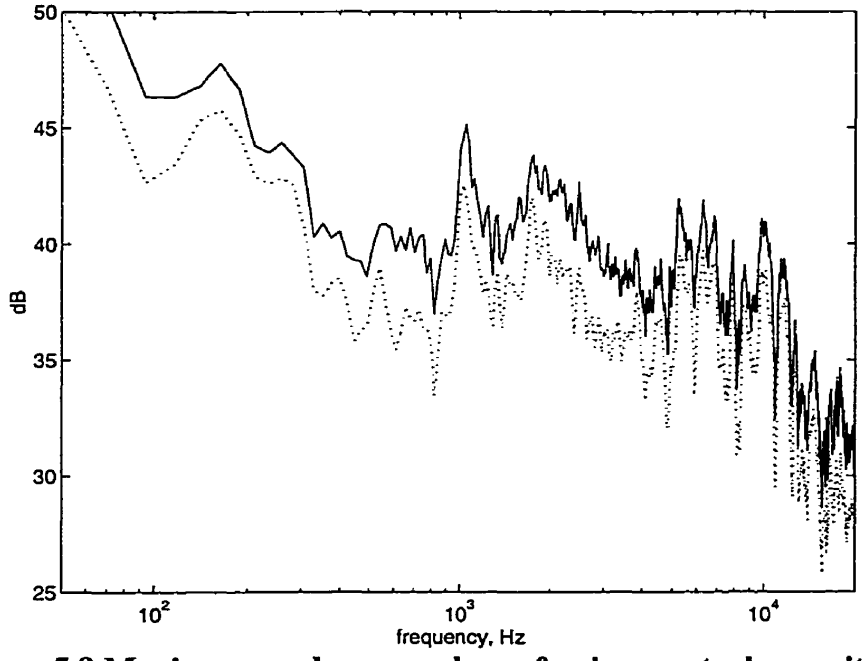


Figure 5.9 Maximum and mean values of noise spectral magnitudes.

One method for overcoming the problem of incomplete noise cancellation employs the overestimation parameter β first proposed by Berouti *et al.* [17]. Although overestimation was originally intended as a means of reducing musical noise, it is also useful for obtaining more complete cancellation of the interfering noise. As its name suggests, this parameter provides an overestimation of the noise in the spectral subtraction process thus ensuring that a greater portion of the noise is canceled. Higher values of β provide a

greater degree of noise suppression and, given a sufficiently high value, the interfering noise can be completely eliminated. A key advantage of the overestimation parameter is its simplicity and several variations to the basic concept have been proposed.

5.4.9 Overestimation based on the expected value and variance of the noise spectral magnitude

While the use of an overestimation parameter effectively reduces the amount of uncanceled noise, the methods proposed to date are limited in that they do not account for the variations of each spectral magnitude component. That is, they are based solely on the mean (or maximum) of the spectral magnitude of the noise. Therefore, a new variant to the overestimation parameter is proposed which makes use of both the mean and variance of the spectral magnitude of the noise. The following variation to the spectral subtraction equation is proposed,

$$\left|\hat{S}_w(e^{j\omega})\right|^{\gamma} = \left|Y_w(e^{j\omega})\right|^{\alpha} - \left(E\left[\left|N_w(e^{j\omega})\right|^{\alpha}\right] + \beta \cdot STD\left[\left|N_w(e^{j\omega})\right|^{\alpha}\right]\right), \qquad (5.36)$$

where $STD[\cdot]$ represents the standard deviation. Both the mean and standard deviation of the spectral magnitude of the noise are estimated during the time interval just prior to speech activity. This new method for overestimating the noise spectral magnitude provides additional flexibility and informal listening tests indicate that it tends to give better noise cancellation performance. This appears to be due to the fact that it more accurately reflects the variations in the interfering noise.

5.4.10 Timbral Effects and Loss of Signal Components

Due to the same mechanism which can cause some of the noise to go uncanceled, spectral subtraction can also cause part of the desired signal to be removed. When the level of a spectral component of the noise within a given processing frame is lower than the noise estimate, a portion of the signal at that frequency will be canceled. The results of this can be heard as either a change in the timbre of the desired signal or a loss of low level signal components.

One way to limit the amount of desired signal which is canceled in the spectral subtraction process is to underestimate the level of the noise. Of course, this will result in incomplete noise cancellation. Therefore, a balance must be found in order to obtain a sufficient degree of noise cancellation without overly distorting the underlying speech signal. Informal listening tests indicated that the new overestimation method proposed in

the previous section tends to provide a good compromise between these two conflicting requirements.

5.4.11 Phase Distortions

One of the fundamental assumptions of the spectral subtraction process was that the phase of the desired signal could be successfully approximated by the phase of the noisy input signal. This approximation means that there will be some error in the processed signal. Even if one could somehow know the exact spectral magnitude of the noise, spectral subtraction does not offer any mechanism to determine the phase of the noise. There are two main types of audible artifacts which occur as a result of this phase error: roughness and temporal smearing. Roughness is primarily heard during sustained sounds such as vowels in speech. Conversely, temporal smearing is more readily heard as pre-echoes and post-echoes occurring near a transient signal.

Vary [19] showed that there is a direct relation between the expected maximum phase error and the signal-to-noise ratio for complex gaussian noise. Vary derived the following relation

$$E[\Delta\Phi_{\max}] = \sin^{-1}\left(\sqrt{\frac{\pi}{2}} \cdot \frac{N(e^{j\omega})}{S(e^{j\omega})}\right), \tag{5.37}$$

where $\Delta\Phi_{\max}$ is the maximum phase deviation, $N(e^{j\omega})$ is the noise power, and $S(e^{j\omega})$ is the signal power at frequency ω. This expression is plotted in Figure 5.10 as a function of the signal-to-noise ratio.

There is one point on the curve shown in Figure 5.10 which is of particular interest. For a signal-to-noise ratio of about 6 dB, the resulting expected maximum phase error is $\pi/4$. This point is of interest because it has been shown that while the ear is relatively insensitive to phase, the threshold at which a random phase error becomes audible is about $\pi/4$ radians [19]. For larger phase errors the speech takes on a *rough* quality. Therefore, it can be concluded that roughness due to phase error will not be audible provided that the signal-to-noise ratio is above 6 dB.

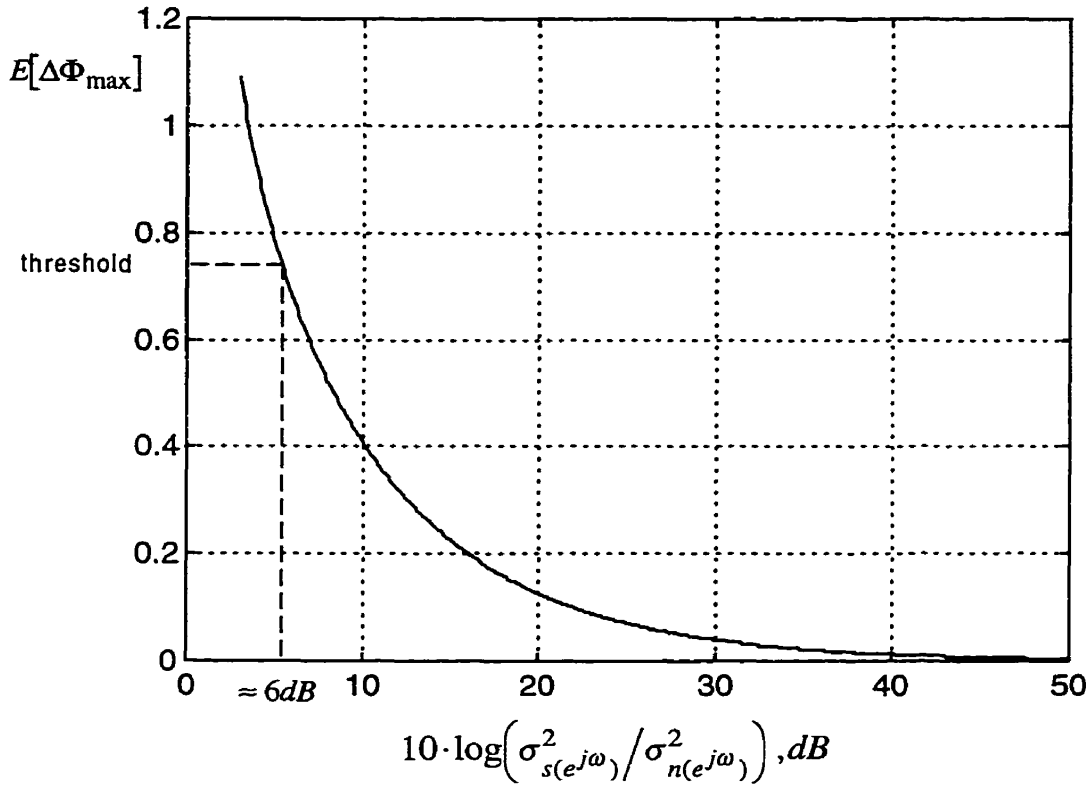


Figure 5.10 Maximum expected phase error due to the addition of gaussian noise as a function of the signal-to-noise ratio.

5.4.12 Time Aliasing and Temporal Smearing

The effects of phase errors of less than $\pi/4$ radians in the spectral subtraction process can also be heard as temporal smearing due to time aliasing. Allen [98] noted that in an analysis-synthesis system, such as the FFT-IFFT process in the spectral subtraction algorithm, any modification to the signal in the spectral domain is equivalent to filtering in the time domain. Since time domain filtering typically results in an increase in the length of a signal, the spectral subtraction process can distort the time waveform due to time aliasing. The temporal aliasing is a result of the circular convolution of the signal with the time domain response (impulse response) of the modification [106].

Time aliasing is illustrated in Figure 5.11. The upper plot shows a 64 sample reference signal which has been augmented to 128 samples through zero padding. The lower plot shows the signal after it has been transformed to the spectral domain, modified, and inverse transformed back to the time domain. It can be seen that due to the spectral modification, samples 65 to 128 in the lower plot are no longer equal to zero. Rather, the

spectral subtraction process caused some leakage of the signal into the zero padded samples.

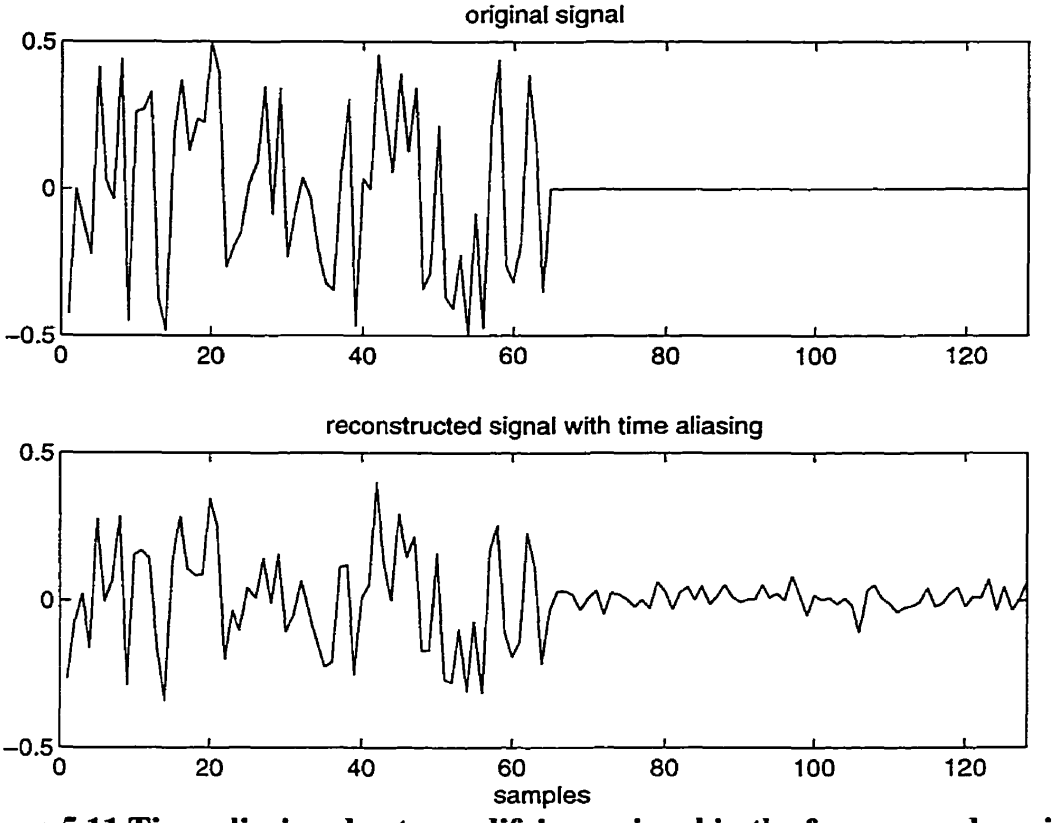


Figure 5.11 Time aliasing due to modifying a signal in the frequency domain.

Another way of viewing time aliasing is as a mismatch between the phase and magnitude of the reconstructed signal. In the spectral subtraction process, an estimate of the spectral magnitude of the noise $E[|N_w(e^{j\omega})|]$ is subtracted from the magnitude of the noisy signal $|Y_w(e^{j\omega})|$. However, in constructing the estimate of the clean signal, the phase of the noisy signal $\Phi_{Y_w}(e^{j\omega})$ is combined with the magnitude of the estimate of the clean signal $|\hat{S}_w(e^{j\omega})|$. This is not the correct phase for the estimated magnitude of the clean signal and so temporal aliasing will occur. In order to avoid temporal aliasing, it would be necessary to have the phase $\Phi_{\hat{S}_w}(e^{j\omega})$ corresponding to the *estimate* of the clean signal. It should be noted that this is not the same as the phase of the actual clean signal. Since this phase information is not available, temporal aliasing results.

In the traditional spectral subtraction process the time samples are not typically zero padded prior to performing the FFT. In this case, the samples which are seen as leakage in the lower plot would be aliased (superimposed) onto the first 64 samples. This results

in a temporal smearing of the signal which can be heard as pre-echoes or post-echoes. If a transient (signal onset) occurs in the middle or latter portion of the analysis window, temporal smearing will occur before the transient thus creating a pre-echo. An example of a pre-echo due to time aliasing is shown in Figure 5.12. where the upper plot shows the reference time waveform without a pre-echo, and the lower plot shows a pre-echo due to time aliasing. The pre-echo begins just before sample 2000 and lasts for about 1500 samples. Note that the pre-echo appears just prior to a sharp transient in the signal. A second pre-echo can be seen near sample 8000 and lasting for just over 1000 samples.

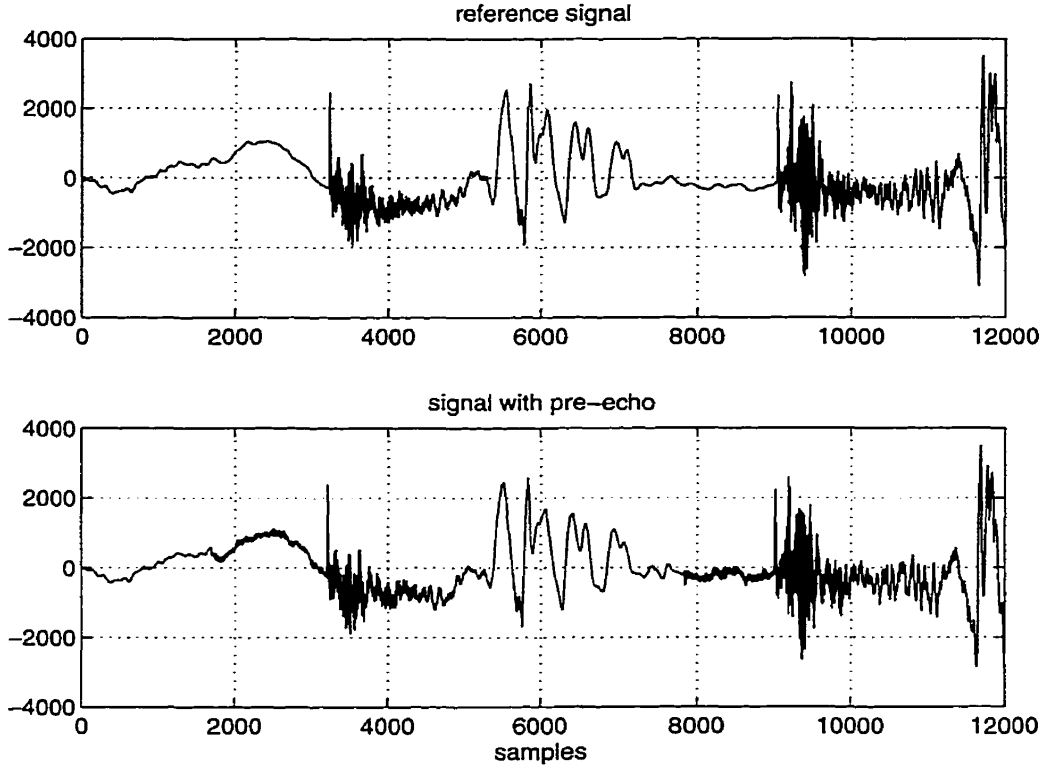


Figure 5.12 Pre-echoes resulting from time aliasing.

Pre-echoes as shown in Figure 5.12 are commonly produced by transform based perceptual audio coders. Subjective tests by Shlien and Soulodre [107] have shown that these pre-echoes can be easily detected by certain groups of listeners and so an effort should be made to minimize them in the spectral subtraction process

5.4.13 Zero Padding with Truncation

One means of addressing the issue of time aliasing is to use zero padding when performing the FFT. However, placing the zero samples at the end of the data samples will create post-echoes which may be audible since there is now a delay between the signal and the

leakage. Placing the zeros before the data samples will result in an exaggerated form of pre-echo. In order to resolve this matter, the following scheme was devised. A frame of M data samples is zero padded prior to performing an FFT of size N ($M<N$). The zeroes are placed after the data samples. Following the spectral subtraction operation, the data is transformed back to the time domain. As seen earlier, this will result in leakage of non-zero samples into the zero padded portion of the time waveform. The length of the time waveform is then truncated and only the first M samples are used to construct the noise reduced signal. The $N-M$ samples which contain the leakage are discarded since they tend to be detrimental to the quality of the processed signal. In the example of Figure 5.11, samples 65 to 128 of the lower plot would be discarded. Informal listening tests indicate that with $N \approx 2M$, this scheme effectively eliminates temporal smearing artifacts (pre-echoes and post-echoes) and provides a distinct improvement in the perceived quality of processed signals containing transients.

5.5 Summary

In this chapter signal enhancement schemes based on estimating the short-time spectral magnitude of the signal were described. This approach is useful for reducing both the periodic component and the cyclical random component of the camera noise. It was seen that the spectral subtraction process can be interpreted as a zero-phase filter, and a generalized form of this zero-phase filter provides a high degree of flexibility and control in the noise suppression process.

The artifacts which are created as a result of the spectral subtraction process were identified and a variety of schemes for limiting their audibility were introduced. A careful balance of these schemes can provide an acceptable degree of noise reduction for moderate levels of camera noise. For higher levels of camera noise however, the traditional spectral subtraction process yields unacceptable levels of distortion and artifacts in the reconstructed signal.

6. SPECTRAL SUBTRACTION USING SUB-FRAMING AND SUBBANDS

In the previous chapter it was seen that the spectral subtraction process can provide a degree of noise reduction but may cause audible artifacts as a result. These artifacts become more pronounced (audible) as the signal-to-noise ratio of the corrupted signal decreases. Informal listening tests have shown that when removing low level camera noise a proper balance between the choice of an appropriate suppression curve, a moderate overestimation factor based on first and second order statistics, and a modified survival algorithm with carefully chosen thresholds may be employed to reduce the artifacts to an acceptable level. However, as the level of the camera noise increases, either the artifacts produced by spectral subtraction cannot be entirely suppressed, or the quality of the underlying speech becomes unacceptably distorted. Therefore, in its present incarnation, spectral subtraction can not be used for removing camera noise when the level of the noise exceeds some value. In this chapter, a new scheme is described which takes advantage of the repetitive nature of camera noise and allows spectral subtraction to be successfully employed under more severe levels of camera noise.

6.1 Spectral Subtraction using Sub-frames

The results of the acoustic measurements described in Chapter 3 showed that camera noise $n(k)$ can be modeled as the sum of a periodic component $p(k)$ and a cyclical random component $c(k)$ in the form,

$$n(k) = p(k) + c(k) \ . \tag{6.1}$$

Both $p(k)$ and $c(k)$ are related to the film rate of the camera (i.e., 24 frames per second). The methods developed in Chapter 4 were intended to address $p(k)$ and did nothing to reduce $c(k)$.

Another way to view the camera noise is as a series of noise bursts which coincide with the film rate. The noise bursts consist of an initial peak pulse containing a large portion of the noise energy followed by an interval of lower level noise. That is,

$$n(k) = n_{peak}(k) + n_{null}(k) \ . \tag{6.2}$$

$p(k)$ and $c(k)$ each contribute to both the peak and null portions of the noise, so that

$$n(k) = p_{peak}(k) + p_{null}(k) + c_{peak}(k) + c_{null}(k) \ . \tag{6.3}$$

The reason for viewing the noise in this manner is that the level of the noise is greater during the "peak" than during the "null" portions of each noise burst. As such, for a given level of speech, the signal-to-noise ratio during the peaks of the noise will be lower than during the nulls. Since the performance of spectral subtraction is directly dependent on the signal-to-noise ratio of the corrupted speech, this method of viewing the noise can be exploited to improve the noise reduction by allowing the processing to be directed by the peak-and-null property of the camera noise. To do this, the spectral subtraction process is divided into *sub-frames*.

The concept of *sub-framing* is best understood with the help of Figure 6.1 which shows a series of camera noise pulses with overlapping (50%) Hanning windows superimposed. The lengths of the windows, and hence the processing frames, are chosen to exactly coincide with the period T of the camera noise. Furthermore, the windows are aligned such that they are centered on either the peak or null portion of a noise pulse. In other words, the windows are aligned to alternately provide a higher noise level followed by a lower level of noise.

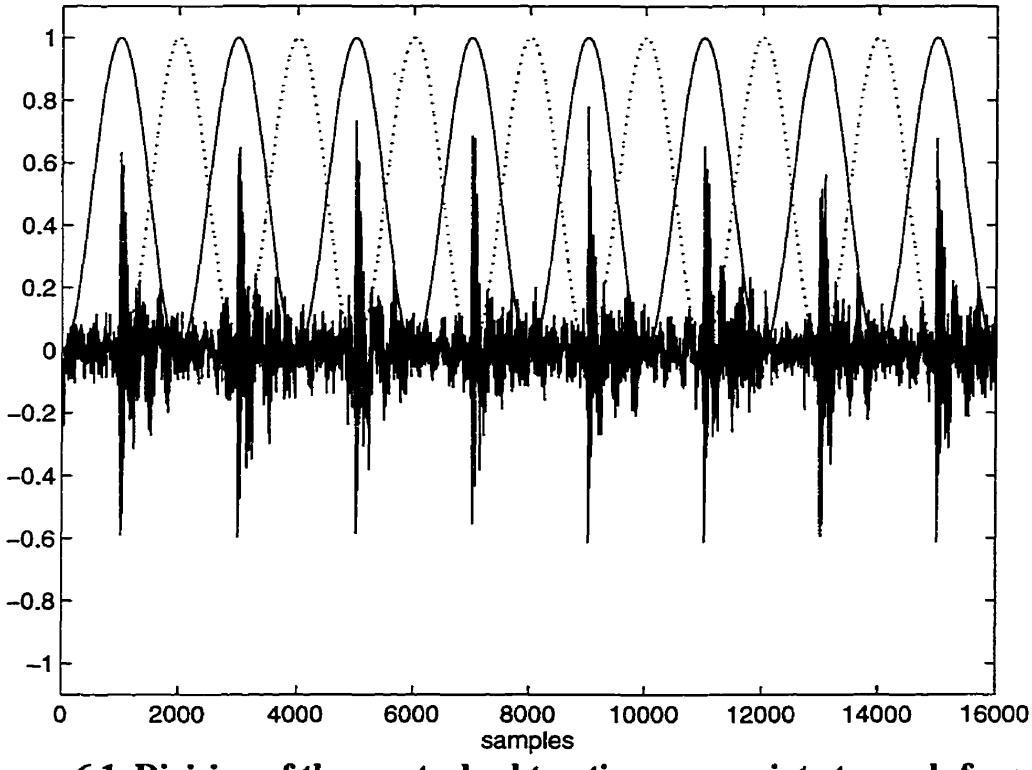


Figure 6.1 Division of the spectral subtraction process into two sub-frames.

In sub-framing we operate two interleaved spectral subtraction processes. One process operates on the peak portions of the noise (solid curve) while the other process operates on the null portions (dotted curve). Let $\Psi_{peak}^{(r)}(\omega_n)$ and $\Psi_{null}^{(r)}(\omega_n)$ be the zero-phase spectral subtraction filters operating on the peak and null portions of the camera noise. The two spectral subtraction processes use separate noise estimates: one based on the level of the noise during the peaks $E[|N_{peak}(e^{j\omega})|^2]$ and one based on an estimate of the noise during the nulls $E[|N_{null}(e^{j\omega})|^2]$. Because of the difference in the level of the noise between the peak and the null portions of the noise pulses, the spectral subtraction process operating on the null portions is operating at a significantly higher signal-to-noise ratio. As a result, more moderate processing can be used and thus the resulting artifacts are far less audible. During the peak portions, the signal-to-noise ratio is lower and more aggressive processing must be applied to the signal. However, this more aggressive processing is done over a shorter length of time and thus the resulting artifacts are relatively short-lived and are consequently less audible. The reduced audibility of the short-lived artifacts is due in part to the masking that occurs in the ear. This will be described further in Chapter 7.

It should be noted that the two interleaved spectral subtraction processes are independent of each other, and thus the parameters of the processes can be individually optimized to provide the best performance. For example, the interleaved processes can use different noise suppression curves, different overestimation factors, and different thresholds for their survival algorithms. The use of sub-frames provides a greater degree of freedom in the spectral subtraction process and inherently reduces all types of audible artifacts.

The above sub-framing process can be described mathematically as follows. Let $Y_{peak}^{(r)}(\omega_n)$ and $Y_{null}^{(r)}(\omega_n)$ be the DFT's of the rth peak and null frames respectively,

$$Y_{peak}^{(r)}(\omega_n) = Y(2rB,n) = \sum_{k=0}^{L-1} y(2rB+k)w(k)e^{-j\frac{2\pi}{N}nk} \tag{6.4}$$

$$Y_{null}^{(r)}(\omega_n) = Y((2r+1)B,n) = \sum_{k=0}^{L-1} y((2r+1)B+k)w(k)e^{-j\frac{2\pi}{N}nk} \tag{6.5}$$

where r and n are integers such that $-\infty < r < \infty$, $0 \le n \le N\text{-}1$, and $N \ge L \ge B$. N is the number of points in the DFT, L is the length of the window $w(k)$, and B is number of

samples by which the window is shifted from frame to frame. Typically, the window $w(k)$ will not be rectangular and there will be some overlapping of the frames.

Given $Y_{peak}^{(r)}(\omega_n)$ and $Y_{null}^{(r)}(\omega_n)$, the estimates of the desired signal during these sub-frames is found using

$$\hat{S}_{peak}^{(r)}(\omega_n) = \Psi_{peak}^{(r)}(\omega_n) \cdot Y_{peak}^{(r)}(\omega_n) \tag{6.6}$$

and

$$\hat{S}_{null}^{(r)}(\omega_n) = \Psi_{null}^{(r)}(\omega_n) \cdot Y_{null}^{(r)}(\omega_n) \ . \tag{6.7}$$

The corresponding time-domain versions of the enhanced signal can be found using the inverse DFT,

$$\hat{s}_{peak}^{(r)}(k) = \frac{1}{N} \sum_{n=0}^{N-1} \hat{S}_{peak}^{(r)}(\omega_n) e^{j\frac{2\pi}{N}nk} \tag{6.8}$$

and

$$\hat{s}_{null}^{(r)}(k) = \frac{1}{N} \sum_{n=0}^{N-1} \hat{S}_{null}^{(r)}(\omega_n) e^{j\frac{2\pi}{N}nk} \tag{6.9}$$

The sub-frame based spectral subtraction process is shown in Figure 6.2

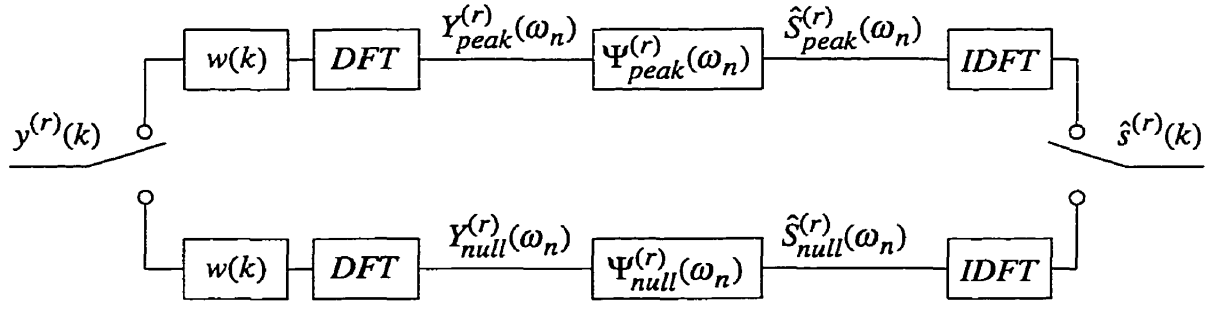


Figure 6.2 Sub-frame based spectral subtraction.

The sub-frame based spectral subtraction process assumes that there is a significant difference in the level of the noise during the peak and null portions of the noise. Figure 6.3 shows an example of the peak and null noise estimates derived from one of the acoustic measurements described in Chapter 3. The upper curve represents the estimate of the peak portion of the camera noise, while the lower curve represents the estimate of the null portion. It can be seen that at some frequencies the difference between the two curves is as much as 10 to 15 dB. This means that, by matching the noise reduction process to the characteristics of the noise, (i.e., using sub-framing) the signal-to-noise ratio

during the null portions is 10 to 15 dB higher at these frequencies than during the peak portions. Therefore, the artifacts created by the spectral subtraction process will be greatly reduced for these sub-frames.

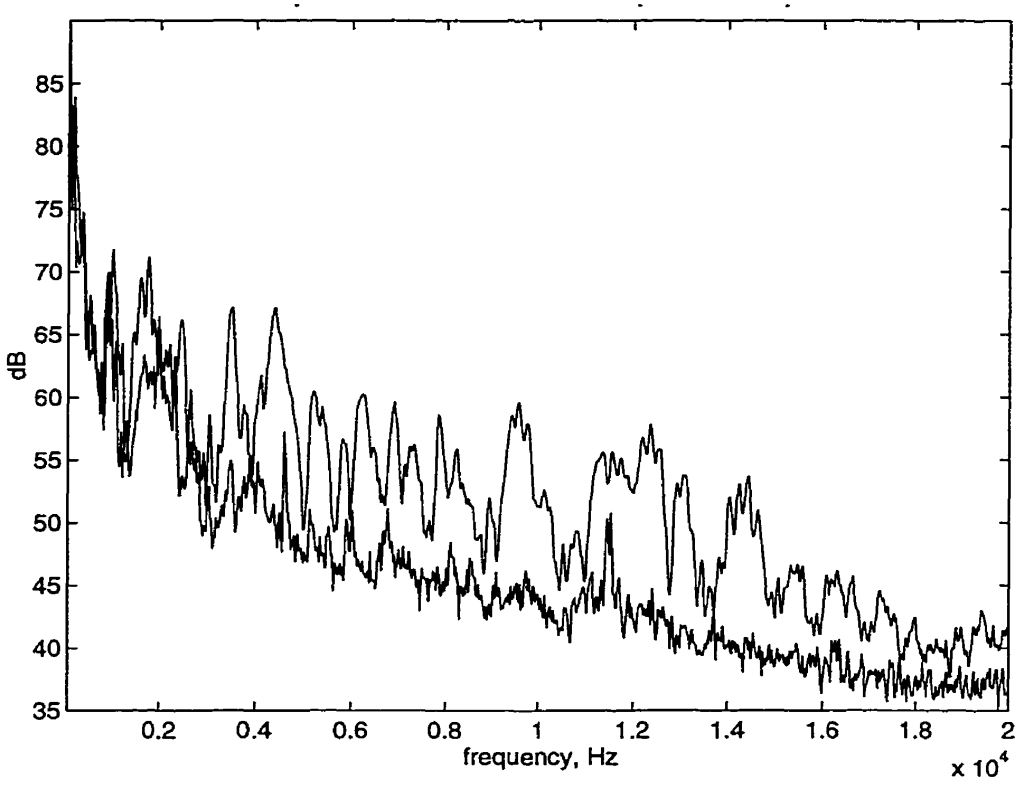


Figure 6.3 Peak and null sub-frame noise estimates.

While the use of sub-frame based spectral subtraction has been proposed in the context of reducing camera noise, one can recognize that this scheme could be used to improve the performance of spectral subtraction algorithms when dealing with other cyclical or repetitive noise sources.

6.1.1 Window Alignment and Frame Synchronization

To avoid modulation of the noise floor of the output signal, the sub-frames of the spectral subtraction process must be synchronized to the film rate (24 frames per second). As noted in Chapter 3, the digital audio recordings of the camera noise were made with a sampling rate, F_S=48000 Hz which is an integer multiple of the film rate. For a sampling rate of 48000 Hz and assuming for the moment that there is no jitter, the pulses of the camera noise occur every 2000 samples. In order to be synchronized to the pulses of the camera noise, the length of the window $w(k)$ used in the sub-framing scheme must also be 2000 samples. That is, we require $L=T$ and $B=T/2$ in equation (6.4). In practice, while

the spectral subtraction algorithm operates on blocks of 2000 data samples, these blocks are zero padded to augment them to a power of 2 to enable the use of a radix-2 FFT [106].

In conjunction with the frame synchronization, it is also necessary to ensure that the windows are properly aligned with the pulses of the camera noise. That is, steps must be taken to align the "peak" window to the peaks of the noise pulses. Improper window alignment will limit the effectiveness of the sub-framing scheme since the difference in the peak and null noise estimates will not be maximized. To examine the importance of frame synchronization and window alignment, it is instructive to consider Figure 6.4 and Figure 6.5

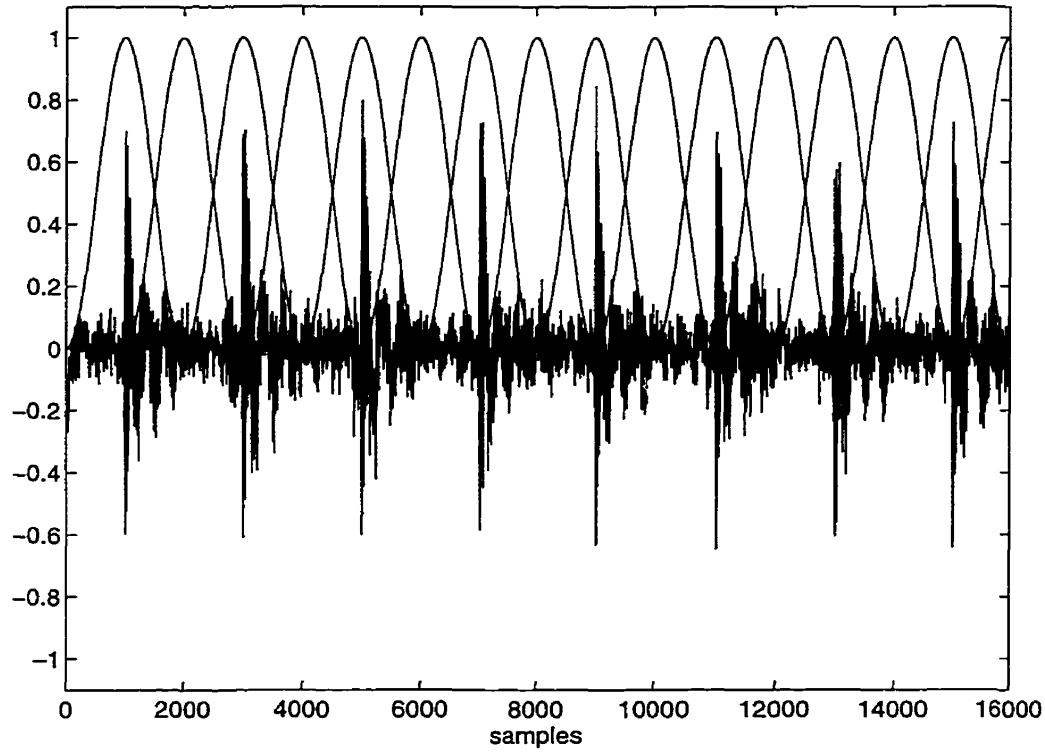


Figure 6.4 Process synchronized to the film rate and windows correctly aligned to the noise pulses.

Figure 6.4 shows the sub-frames correctly synchronized to the film rate (L=T) and the windows properly centered on the peaks of the noise pulses. It can be seen that alignment and synchronization are maintained over all of the noise pulses. Conversely, Figure 6.5 represents a system which is not synchronized to the film rate (L>T). Here, it can be seen that, while the first window is correctly aligned to the noise pulse, subsequent windows quickly drift out of alignment with the noise pulses. In this example, the size of the sub-

frame was 2048 samples which was chosen to represent an appropriate selection for the FFT and inverse-FFT operations.

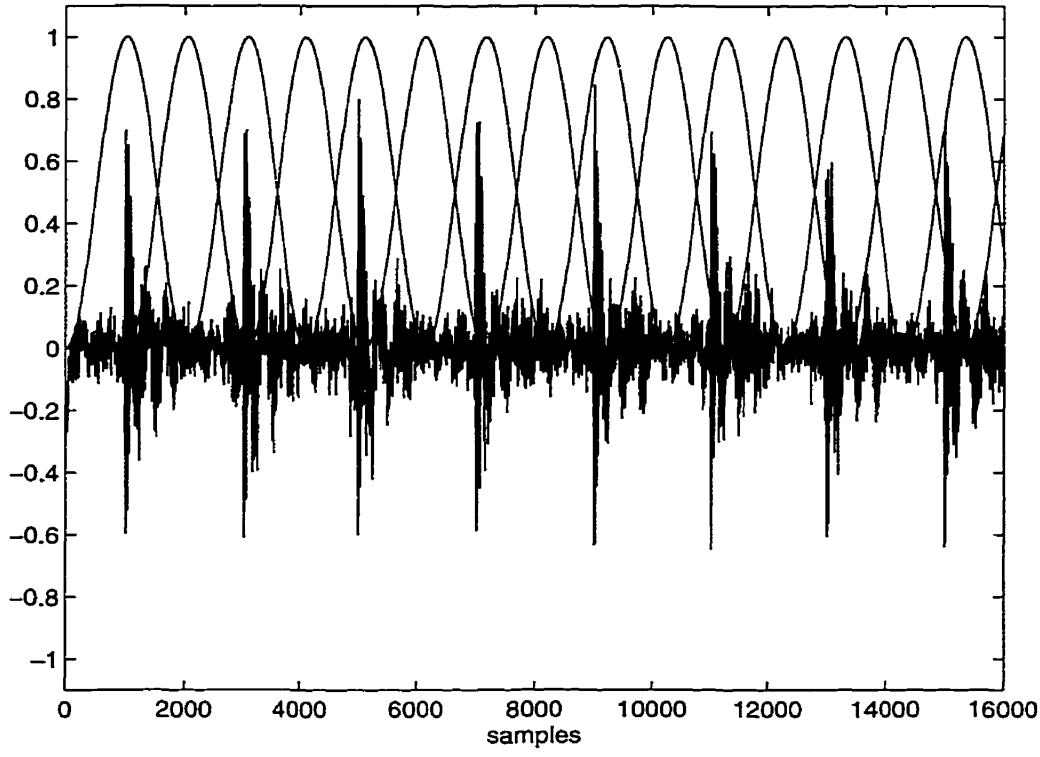


Figure 6.5 Process not synchronized to the film rate.

The effects of non-synchronized sub-frames and improper window alignment can be seen more readily in Figure 6.6 and Figure 6.7. Figure 6.6 shows a spectrogram for the case where the sub-frames are synchronized to the film rate and the windows are properly aligned. The spectrogram consists of an alternating sequence of dark and light vertical bands. These correspond respectively to the peaks and nulls of the camera noise. The vertical bands remain consistent and well defined over the entire spectrogram. Figure 6.7 shows a spectrogram of the camera noise with a sub-frame size of 2048 samples, and thus the process is not synchronized to the film rate. At the start of the process, the light and dark vertical bands representing the nulls and peaks are well defined. However, the pulses of the noise soon begin to drift with respect to the windows and at about 500 ms the distinction between noise peaks and nulls is lost. The benefits of sub-framing are defeated during this segment of the signal and very poor noise reduction would occur. The pulses continue to drift until they are once again temporarily aligned with the processing at about 800 ms. Thus it can be seen that the performance of the spectral subtraction process would modulate over time.

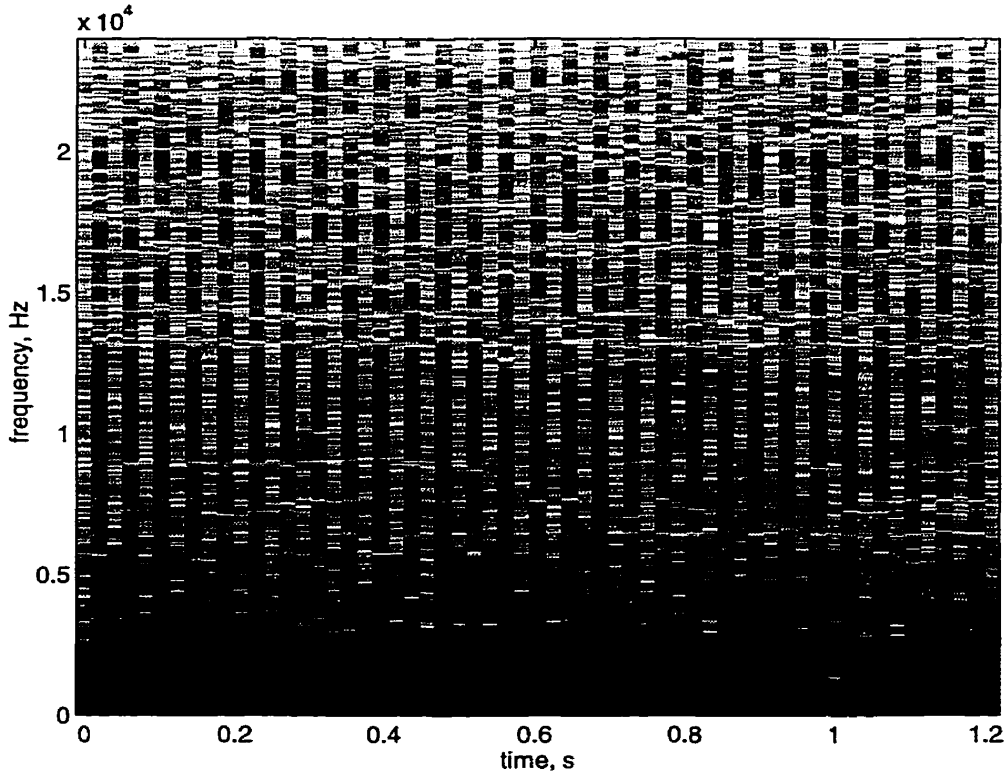


Figure 6.6 Spectrogram of camera noise with process synchronized and aligned to the camera noise.

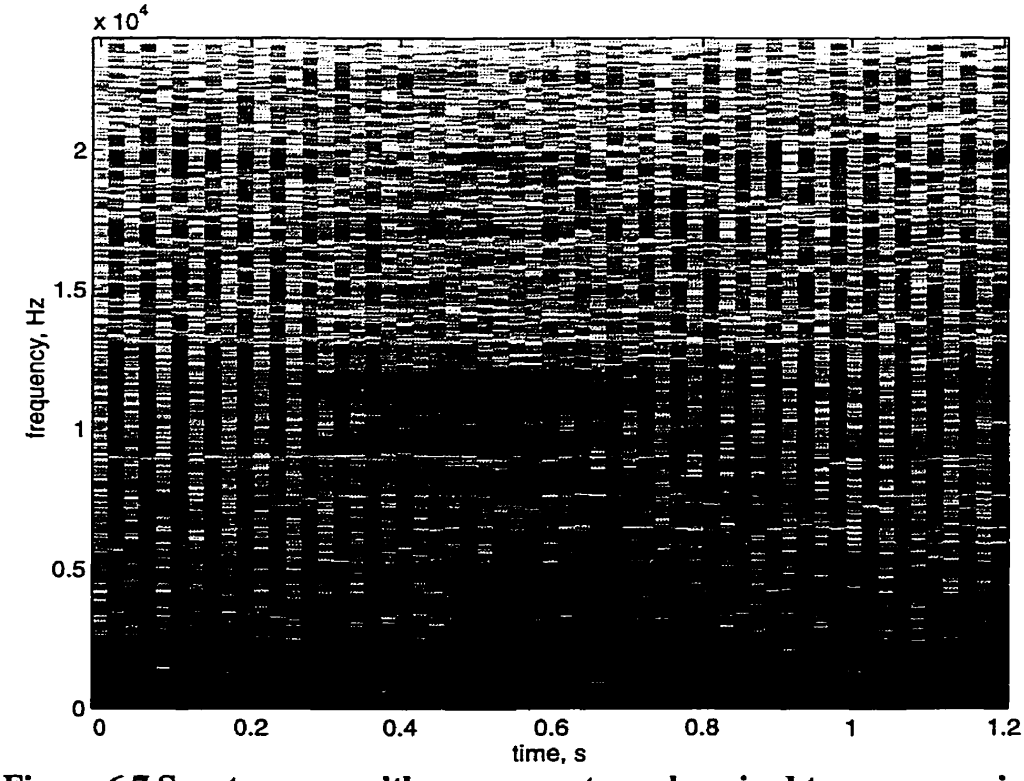


Figure 6.7 Spectrogram with process not synchronized to camera noise.

The above discussion assumed that there was no jitter in the periodic component $p(k)$ of the camera noise. In Chapter 4 it was seen that a significant amount of jitter can occur and so its impact on the performance of the sub-frame spectral subtraction process must be considered. First it should be realized that the jitter caused the timing of $p(k)$ to vary in both the positive and negative directions. To reflect this, the model for the periodic component of the camera noise was modified to include the zero-mean random variable $\xi(k)$,

$$p(k)=\sum_{l=-\infty}^{\infty} Q(k+lT+\xi(k)) \; . \tag{6.10}$$

Therefore, even with the jitter, the noise reduction process will remain synchronized on average over time assuming $L=T$ and $B=T/2$. Moreover, the spectral subtraction only processes the magnitude of the input signal and does not alter the phase. Since a slight shift (due to jitter) between the signal and the processing window will have only a minor effect on the spectral magnitude of the signal, we can conclude that explicit compensation for the jitter is not necessary.

6.1.2 A Simple Method for Frame Synchronization and Window Alignment

It was seen in the previous section that, in order for sub-framing to operate successfully, the sub-frames must be synchronized to the film rate and the windows must be aligned to the noise pulses. In this section a simple method is described for ensuring frame synchronization and window alignment.

Because the camera noise is mixed with the desired speech signal, the pulses of the noise may not be easily detected directly from the time waveform. Similarly, a spectrogram of the corrupted signal may not be sufficient for determining the precise locations of the camera noise pulses. This is particularly true for high signal-to-noise ratios where the level of the noise is low relative to the speech signal. However, by taking an approximation to the second derivative of the corrupted signal, the peak pulses of the camera noise become readily apparent. This is due to the steep slopes associated with the onset of each noise pulse. This method was proposed by Kasparis and Lane [108] as a means of detecting scratches on vinyl records. The process is shown in Figure 6.8 and can be described by the following equation,

$$u(k)=|y(k)-2\cdot y(k-1)+y(k-2)|. \tag{6.11}$$

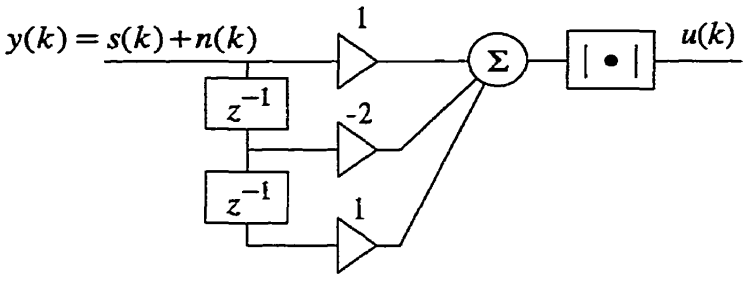


Figure 6.8 Circuit used to detect peaks of the noise pulses.

The result of this process can be seen in Figure 6.9. The upper plot shows the time waveform of the noisy input signal $y(k)$. The camera noise can be seen as a series of pulses between samples 30000 and 50000, but is hidden at other points in the waveform by the speech signal. The lower plot of the figure shows the output $u(k)$ of the circuit when the waveform shown in the upper plot is presented at the input. The location of each noise pulse is easily seen in this plot. With this simple method, the sub-frames of the two spectral subtraction processes can be easily synchronized to the film rate of the camera, and the processing windows can be aligned to the noise pulses.

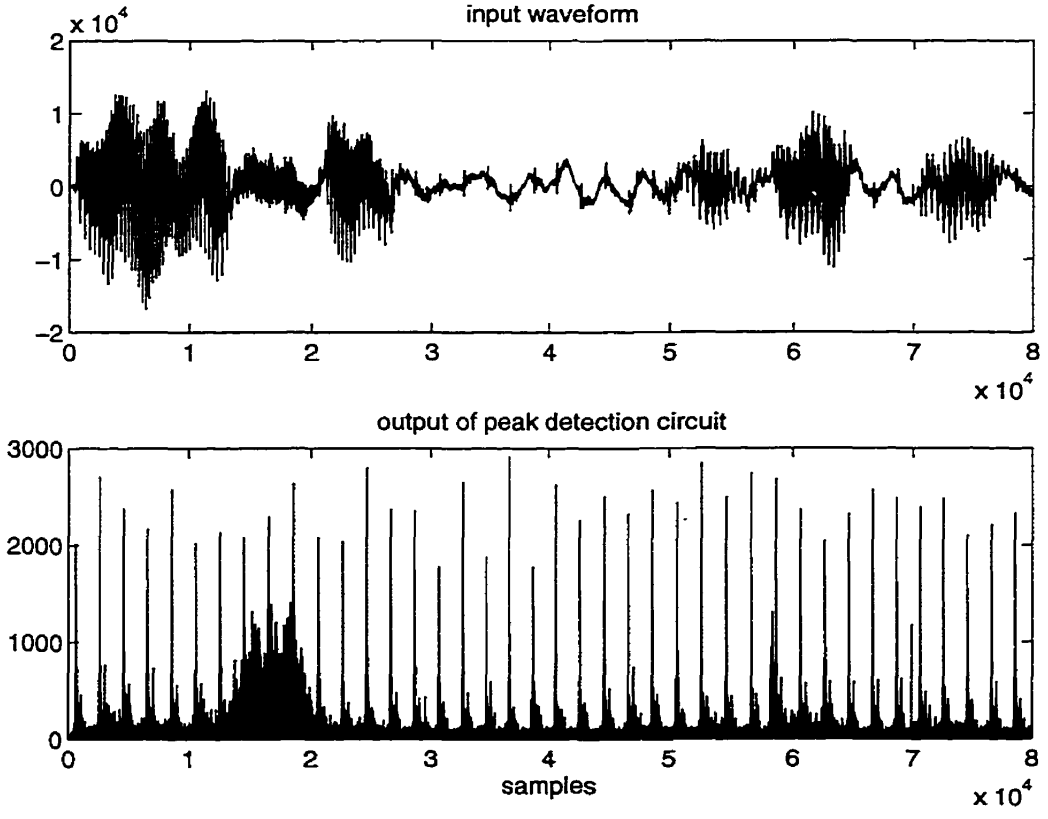


Figure 6.9 Input and output of noise peak detector.

6.1.3 Multi Sub-Framed Spectral Subtraction

The concept of sub-framed spectral subtraction was introduced using 2 sub-frames. It may be desirable in some instances to divide the signal into multiple sub-frames. Equations (6.4) and (6.5) can be readily modified to accommodate multiple sub-frames,

$$Y_d^{(r)}(\omega_n) = \sum_{k=0}^{L-1} y((Gr+d)B+k)w(k)e^{-j\frac{2\pi}{N}nk} \tag{6.12}$$

where L is the length of the windows, d=0,1,...,G-1 is the index for the various sub-frames, G is the number of sub-frames, and B is the number of samples by which the windows overlap. The sum of the lengths of the sub-frames must equal the period T of the camera noise in order to obtain the benefit of sub-framed spectral subtraction.

$$G(L-B) = T \tag{6.13}$$

The concept of sub-framing can be further generalized by allowing sub-frames to be of different lengths. Again, the sum of the lengths of the sub-frames must be equal to the period T of the camera noise. The benefit of non-uniform sub-frames will be seen in a later section when they are used in conjunction with subband filtering.

A logical question to consider is whether or not further benefit can be gained by using more than 2 sub-frames. This matter was considered and spectral subtraction algorithms were implemented using 2, 4, and 8 sub-frames. Thus the sizes of the sub-frames were 2000, 1000, and 500 samples respectively.

An analysis of the performance of spectral subtraction using these various numbers of sub-frames revealed that increasing the number of sub-frames does not necessarily result in improved performance. This is because as the number of sub-frames increases there are fewer data samples within each sub-frame and consequently there is greater variation in the average magnitude spectrum of the noise from frame to frame. This is particularly true at low frequencies. Furthermore, it is not evident that the makeup of the camera noise at lower frequencies warrants more than 2 sub-frames. On the other hand, using more (i.e., shorter) sub-frames may provide better performance at higher frequencies where the duration of the camera noise is shorter. That is, at lower frequencies a spectral subtraction process based on 2 sub-frames is well suited to the characteristics of the noise, while higher frequencies are better matched to a 4 or 8 sub-frame process.

6.2 *Spectral Subtraction using Subband Filtering*

In this section a method based on decomposing the signal into frequency subbands is described. Noise reduction on the individual frequency subbands is then achieved using separate spectral subtraction processes. The full benefit of the subband processing will be realized in the next section when it is combined with sub-framing.

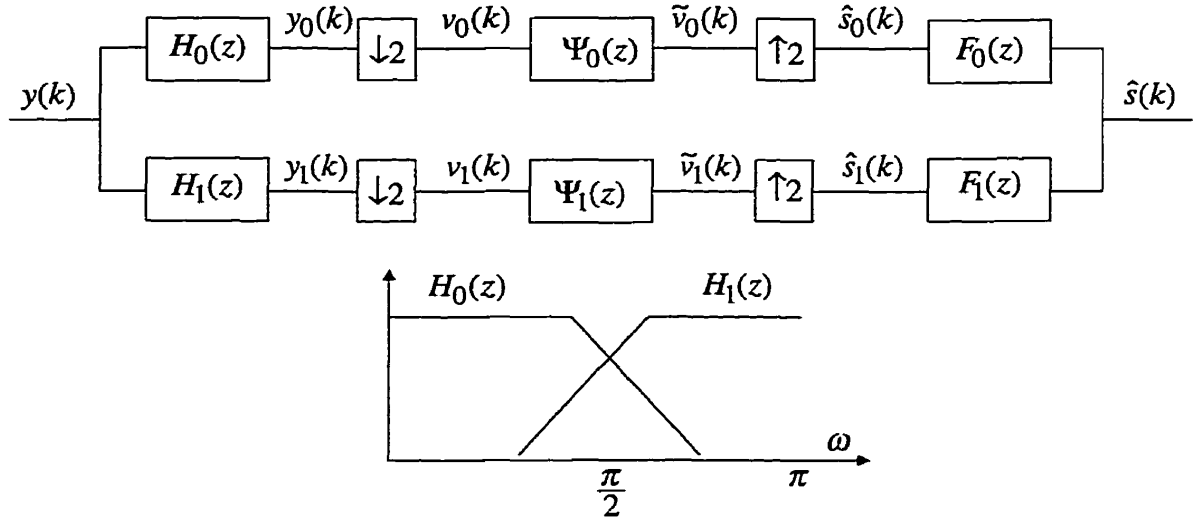


Figure 6.10 Noise reduction based on a 2-subband QMF analysis/synthesis filter bank.

To divide the noisy signal $y(k)$ into frequency bands, a quadrature mirror filter (QMF) bank is employed [109,111,112]. The basic QMF structure is depicted in Figure 6.10. As shown in the figure, $y(k)$ is filtered by a lowpass filter $H_0(z)$ and a highpass filter $H_1(z)$.

$$Y_l(z) = H_l(z)Y(z) \quad l = 0,1. \ , \tag{6.14}$$

where $Y(z)$ is the z-transform of $y(k)$.

The cutoff frequencies of the two filters are both set to $\pi/2$. As a result, the subband signals $y_0(k)$ and $y_1(k)$ are each roughly limited to a bandwidth of $\pi/2$. These signals are subsequently decimated by a factor of 2 yielding the critically sampled signals $v_0(k)$ and $v_1(k)$. Using the following relation,

$$V(z^M) = \frac{1}{M}\sum_{q=0}^{M-1} Y(zW_M^q), \tag{6.15}$$

where M is the decimation factor and W_M=exp(-$j2\pi/M$), we can derive an expression for the critically sampled subband signals $V_l(z)$ l=0,1

$$V_l(z)=\tfrac{1}{2}[Y_l(z^{1/2})+Y_l(-z^{1/2})] \tag{6.16}$$

$$=\tfrac{1}{2}[Y(z^{1/2})H_l(z^{1/2})+Y(-z^{1/2})H_l(-z^{1/2})]\ . \tag{6.17}$$

The second term in (6.17) represents the aliasing that occurs as a result of the decimation.

The signals $v_0(k)$ and $v_1(k)$ are then processed by two separate noise reduction algorithms Ψ_l l=0,1. Since the processes are separate, they can operate using different numbers of sub-frames. The enhanced subband signals $\tilde{v}_0(k)$ and $\tilde{v}_1(k)$ are expanded (upsampled) by a factor of 2 to give $\hat{s}_0(k)$ and $\hat{s}_1(k)$ which are filtered through $F_0(z)$ and $F_1(z)$ respectively and summed to produce $\hat{s}(k)$

$$\hat{S}(z)=F_0(z)\hat{S}_0(z)+F_1(z)\hat{S}_1(z)\ . \tag{6.18}$$

The filters $H_0(z)$ and $H_1(z)$ constitute the analysis filter bank while $F_0(z)$ and $F_1(z)$ form the synthesis or reconstruction bank. The analysis/synthesis process can be expressed in matrix-vector form. For the analysis bank we have,

$$\begin{bmatrix}V_0(z)\\V_1(z)\end{bmatrix}=\frac{1}{2}\begin{bmatrix}H_0(z^{1/2}) & H_0(-z^{1/2})\\H_1(z^{1/2}) & H_1(-z^{1/2})\end{bmatrix}\begin{bmatrix}Y(z^{1/2})\\Y(-z^{1/2})\end{bmatrix} \tag{6.19}$$

and for the synthesis bank,

$$\hat{S}(z)=\begin{bmatrix}F_0(z) & F_1(z)\end{bmatrix}\cdot\frac{1}{2}\begin{bmatrix}V_0(z^2)\\V_1(z^2)\end{bmatrix}. \tag{6.20}$$

By combining (6.19) and (6.20) we get,

$$\hat{S}(z)=\begin{bmatrix}F_0(z) & F_1(z)\end{bmatrix}\cdot\frac{1}{2}\begin{bmatrix}H_0(z) & H_0(-z)\\H_1(z) & H_1(-z)\end{bmatrix}\begin{bmatrix}Y(z)\\Y(-z)\end{bmatrix}. \tag{6.21}$$

Let

$$G_0(z)=\tfrac{1}{2}[F_0(z)H_0(z)+F_1(z)H_1(z)] \tag{6.22}$$

and

$$G_1(z)=\tfrac{1}{2}[F_0(z)H_0(-z)+F_1(z)H_1(-z)]. \tag{6.23}$$

Equation (6.21) can therefore be expressed as

$$\hat{S}(z)=G_0(z)Y(z)+G_1(z)Y(-z) \tag{6.24}$$

where the function $G_0(z)$ describes the transfer characteristics of the filter bank and $G_1(z)$ represents the aliasing. In the absence of any noise reduction processing, it is desirable to design $H_0(z)$, $H_1(z)$, $F_0(z)$, and $F_1(z)$ such that $\hat{s}(k)$=$y(k)$. However, due to aliasing, am-

plitude distortion, and phase distortion, $\hat{s}(k)$ may not be identical to $y(k)$. In designing the analysis and synthesis filter banks, tradeoffs must be made between the accuracy with which $\hat{s}(k)$ approximates $y(k)$ and the quality (transition bandwidth, stopband attenuation, filter order, etc.) of the filters.

It is possible to design perfect reconstruction filter banks in which the output signal $\hat{s}(k)$ is identical to the input $y(k)$. That is, $\hat{s}(k)$ is merely a delayed version of $y(k)$,

$$\hat{s}(k) = z^{-n} y(k) \ . \tag{6.25}$$

To examine the design of a perfect reconstruction filter bank, we begin by deriving an expression for $\hat{S}(-z)$ based on equation (6.21)

$$\hat{S}(-z) = \begin{bmatrix} F_0(-z) & F_1(-z) \end{bmatrix} \cdot \frac{1}{2} \begin{bmatrix} H_0(-z) & H_0(z) \\ H_1(-z) & H_1(z) \end{bmatrix} \begin{bmatrix} Y(-z) \\ Y(z) \end{bmatrix} . \tag{6.26}$$

Equations (6.21) and (6.26) can be combined as

$$\begin{bmatrix} \hat{S}(z) \\ \hat{S}(-z) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} F_0(z) & F_1(z) \\ F_0(-z) & F_1(-z) \end{bmatrix} \begin{bmatrix} H_0(z) & H_0(-z) \\ H_1(z) & H_1(-z) \end{bmatrix} \begin{bmatrix} Y(z) \\ Y(-z) \end{bmatrix}, \tag{6.27}$$

or equivalently as

$$\hat{\mathbf{s}}(z) = \frac{1}{2} \mathbf{F}(z) \mathbf{H}^T(z) \mathbf{y}(z) \ . \tag{6.28}$$

Perfect reconstruction requires that

$$\begin{bmatrix} \hat{S}(z) \\ \hat{S}(-z) \end{bmatrix} = \begin{bmatrix} z^{-n} & 0 \\ 0 & (-z)^{-n} \end{bmatrix} \begin{bmatrix} Y(z) \\ Y(-z) \end{bmatrix} \tag{6.29}$$

or equivalently

$$\frac{1}{2} \mathbf{F}(z) \mathbf{H}^T(z) = z^{-n} \begin{bmatrix} 1 & 0 \\ 0 & (-1)^{-n} \end{bmatrix} \tag{6.30}$$

which gives

$$\mathbf{F}(z) = 2z^{-n} \begin{bmatrix} 1 & 0 \\ 0 & (-1)^{-n} \end{bmatrix} (\mathbf{H}^T(z))^{-1} \tag{6.31}$$

$$= \frac{2z^{-n}}{\det \mathbf{H}(z)} \begin{bmatrix} H_1(-z) & -H_0(-z) \\ H_1(z) & -H_0(z) \end{bmatrix}, \tag{6.32}$$

where n is an odd integer. Given that the analysis filter bank has been determined, the synthesis bank which will yield perfect reconstruction is derived from (6.32).

While the above design method is certainly valid, it imposes significant limitations on the possible choice of filter transfer functions. Therefore, we often relax the perfect reconstruction requirement of the filter bank in order to obtain better quality (transition bandwidth, stopband attenuation, etc.) filters. Typically the requirement is relaxed by allowing some (inaudible) amplitude distortion while requiring an alias-canceling filter bank with linear phase response.

To examine the requirements for an alias-canceling filter bank, recall that $G_1(z)$ of equation (6.24) represents the aliased component of the reconstructed signal $\hat{s}(k)$. To eliminate any aliasing, we require $G_1(z) = 0$, or

$$\frac{1}{2}[F_0(z)H_0(-z) + F_1(z)H_1(-z)] = 0. \tag{6.33}$$

This can be easily achieved by applying the following constraints,

$$F_0(z) = H_1(-z) \tag{6.34}$$

and

$$F_1(z) = -H_0(-z). \tag{6.35}$$

Therefore, given $H_0(z)$ and $H_1(z)$, the filter bank will be alias-canceling if $F_0(z)$ and $F_1(z)$ are specified according to (6.34) and (6.35). In many QMF banks, the design is further simplified by having the analysis filters be mirror images of each other. That is,

$$H_1(z) = H_0(-z). \tag{6.36}$$

Given the constraints defined by equations (6.34), (6.35), and (6.36), it can be seen that only one *prototype* filter needs to be designed since the remaining filters are completely specified in terms of $H_0(z)$. To eliminate phase distortions $H_0(z)$ is chosen to be a linear phase FIR filter.

Based on the above, Johnston proposed a design optimization technique which minimizes the amplitude distortion while simultaneously maximizing the stopband attenuation of an alias-canceling linear phase filter bank. Johnston provided a family of filter designs with varying characteristics [110,109]. An important feature of these filters is that they provide better filter characteristics (i.e., stopband attenuation, transition bandwidth) than a perfect reconstruction filter bank of the same order while maintaining the amplitude distortion below an audible level [111].

The filter banks described so far divide the input signal $y(k)$ into two subbands. In the present application it is desirable to filter $y(k)$ into more subbands to allow greater flexibility in selecting the size of the sub-frames across frequencies. It is a straightforward process to extend a two channel QMF design into a filter bank with m channels. Figure

6.11 shows a non-uniform analysis filter bank which divides the input signal into 4 subbands as well as the resulting octave spaced filter responses. It is of course, possible to design a uniform filter bank, but the non-uniform design is better suited to the characteristics of the camera noise.

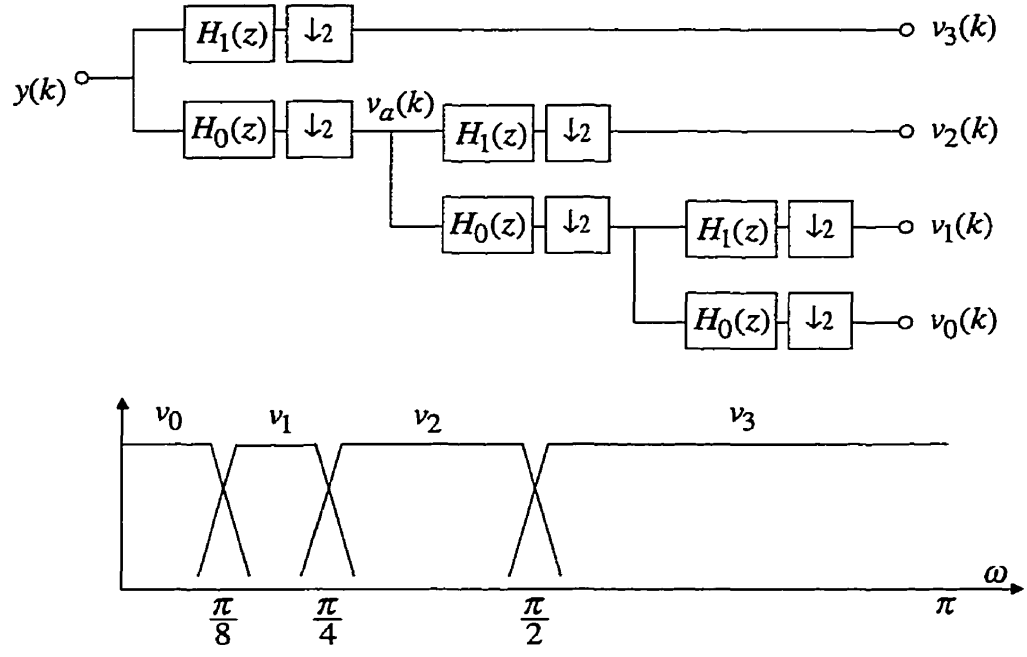


Figure 6.11 Non-uniform QMF analysis bank based on a 3-level binary tree.

The 4-channel filter bank is based on a 3-level binary tree structure wherein the two filters, $H_0(z)$ and $H_1(z)$ are used repeatedly. Therefore, a wide range of filter banks can be devised which are based on the design of a single prototype filter $H_0(z)$. For example, some or all of the subband signals $v_l(k)$ l=0,1,2,3 could be further decomposed using m-channel uniform filter banks based on $H_0(z)$ and $H_1(z)$.

To examine mathematically how the subband signals are derived as a result of the tree structure, consider the subband signal $v_2(k)$ which is derived from $v_a(k)$. As seen earlier,

$$V_a(z) = \tfrac{1}{2}\left[Y(z^{1/2})H_0(z^{1/2}) + Y(-z^{1/2})H_0(-z^{1/2})\right] \tag{6.37}$$

$$V_2(z) = \tfrac{1}{2}\left[V_a(z^{1/2})H_0(z^{1/2}) + V_a(-z^{1/2})H_0(-z^{1/2})\right]. \tag{6.38}$$

Substituting (6.37) into (6.38),

$$\begin{aligned} V_2(z) = {} & \tfrac{1}{2}\left[Y(z^{1/2})H_0(z^{1/2}) + Y(-z^{1/2})H_0(-z^{1/2})\right]H_0(z^{1/2}) \\ & + \tfrac{1}{2}\left[Y(z^{1/2})H_0(z^{1/2}) + Y(-z^{1/2})H_0(-z^{1/2})\right]H_0(-z^{1/2}) \end{aligned} \tag{6.39}$$

Or equivalently, in matrix-vector form,

$$V_2(z) = \frac{1}{4}\begin{bmatrix} H_0(z^{1/2}) & H_0(-z^{1/2}) \end{bmatrix}\begin{bmatrix} H_0(z^{1/2}) & H_0(-z^{1/2}) \\ H_0(z^{1/2}) & H_0(-z^{1/2}) \end{bmatrix}\begin{bmatrix} Y(z^{1/2}) \\ Y(-z^{1/2}) \end{bmatrix} \tag{6.40}$$

$$= \frac{1}{4}\begin{bmatrix} H_0(z^{1/4}) + H_0(-z^{1/4}) & H_0(-z^{1/4}) + H_0(z^{1/4}) \end{bmatrix}\begin{bmatrix} Y(z^{1/2}) \\ Y(-z^{1/2}) \end{bmatrix}. \tag{6.41}$$

The other subband signals can be obtained in a similar fashion.

The corresponding 4-channel synthesis filter bank is shown in Figure 6.12. It can be seen that it consists of a complementary tree structure based on the two filters $F_0(z)$ and $F_1(z)$. It should be noted that, in order for the QMF analysis/synthesis banks to work properly with causal filters, appropriate delays need to be added to the paths of those subbands which are not fully decomposed.

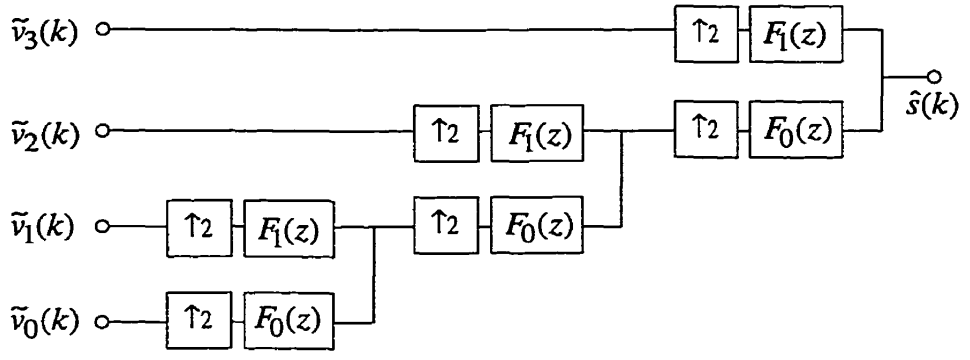


Figure 6.12 Non-uniform QMF synthesis bank based on a 3-level binary tree.

6.2.1 Effect of Processing

The design of the prototype filter is very important to the performance of the subband based noise reduction scheme. As stated earlier, the filter banks used in the present study were alias-canceling and linear phase. However, the alias-canceling property of the design assumes that no processing is done to the subband signals between the analysis and synthesis banks. That is, it is assumed that

$$\tilde{v}_l(k) = v_l(k) \quad \forall l \tag{6.42}$$

where $\tilde{v}_l(k)$ l=0,1,...,m-1 are the m enhanced subband signals. With the noise reduction processing defined as,

$$\tilde{V}_l(\omega_n) = \Psi_l(\omega_n)V_l(\omega_n), \tag{6.43}$$

this is equivalent to assuming that

$$|\Psi_l(\omega_n)| = 1 \quad \forall l, \omega_n \tag{6.44}$$

where $\Psi_l(\omega_n)$ is the zero-phase spectral subtraction filter operating on the lth subband. Whenever the noise reduction algorithm is being applied, we must expect a certain amount of aliasing to occur in the reconstructed signal $\hat{s}(k)$.

To examine how aliasing is affected by the spectral subtraction process, recall the expression for the reconstructed signal

$$\hat{S}(z)=\frac{1}{2}\begin{bmatrix}F_0(z) & F_1(z)\end{bmatrix}\begin{bmatrix}H_0(z) & H_0(-z)\\ H_1(z) & H_1(-z)\end{bmatrix}\begin{bmatrix}Y(z)\\ Y(-z)\end{bmatrix} \tag{6.45}$$

where it was assumed that $\tilde{v}_l(k)=v_l(k)$. Using (6.43) the effect of the spectral subtraction process $\Psi_l(z)$ can be included,

$$\begin{bmatrix}\tilde{V}_0(z)\\ \tilde{V}_1(z)\end{bmatrix}=\begin{bmatrix}\Psi_0(z) & 0\\ 0 & \Psi_1(z)\end{bmatrix}\begin{bmatrix}V_0(z)\\ V_1(z)\end{bmatrix} \tag{6.46}$$

$$\hat{S}(z)=\frac{1}{2}\begin{bmatrix}F_0(z) & F_1(z)\end{bmatrix}\begin{bmatrix}\Psi_0(z_n) & 0\\ 0 & \Psi_1(z)\end{bmatrix}\begin{bmatrix}H_0(z) & H_0(-z)\\ H_1(z) & H_1(-z)\end{bmatrix}\begin{bmatrix}Y(z)\\ Y(-z)\end{bmatrix} \tag{6.47}$$

$$\begin{aligned}\hat{S}(z)=&\frac{1}{2}F_0(z)\Psi_0(z)\left[H_0(z)Y(z)+H_0(-z)Y(-z)\right]\\ &+\frac{1}{2}F_1(z)\Psi_1(z)\left[H_1(z)Y(z)+H_1(-z)Y(-z)\right]\end{aligned} \tag{6.48}$$

Recalling (6.24)

$$\hat{S}(z)=G_0(z)Y(z)+G_1(z)Y(-z)\ , \tag{6.49}$$

$G_0(z)$ and $G_1(z)$ are now defined as

$$G_0(z)=\frac{1}{2}[F_0(z)\Psi_0(z)H_0(z)+F_1(z)\Psi_1(z)H_1(z)] \tag{6.50}$$

and

$$G_1(z)=\frac{1}{2}[F_0(z)\Psi_0(z)H_0(-z)+F_1(z)\Psi_1(z)H_1(-z)]. \tag{6.51}$$

where equation (6.51) represents the aliasing component of $\hat{S}(z)$. Substituting the constraints defined in (6.33) and (6.34)

$$G_1(z)=\frac{1}{2}[H_1(-z)\Psi_0(z)H_0(-z)-H_0(-z)\Psi_1(z)H_1(-z)]. \tag{6.52}$$

$$=\frac{1}{2}H_0(-z)H_1(-z)[\Psi_0(z)-\Psi_1(z)]. \tag{6.53}$$

From (6.53) it can be seen that the amount of aliasing that occurs is related to the difference in the processing $\Psi_0(z)$ and $\Psi_1(z)$ applied to the subbands in the region of overlap. One way of reducing the amount of aliasing is by limiting the difference in the processing applied to the subbands. The amount of aliasing is also related to the transfer

functions of the filters $H_0(z)$ and $H_1(z)$, and so it can also be reduced by careful selection of the filter characteristics. For example, a prototype filter with a narrow transition band and a large stopband attenuation can be used.

Due to the spectral subtraction process, which deliberately alters the input signal, it is highly likely that residual aliasing will occur (i.e. aliased components will not be fully canceled in the QMF synthesis bank). This is particularly true when the input signal $y(k)$ has a low signal-to-noise ratio and more aggressive noise suppression must be applied. In order for the subband processing to be useful, steps must be taken to reduce the aliasing. Therefore, the choice of filters becomes quite important when trying to minimize the creation of any audible artifacts due to aliasing. Comprehensive discussions regarding quadrature mirror filters may be found in [109,111,112].

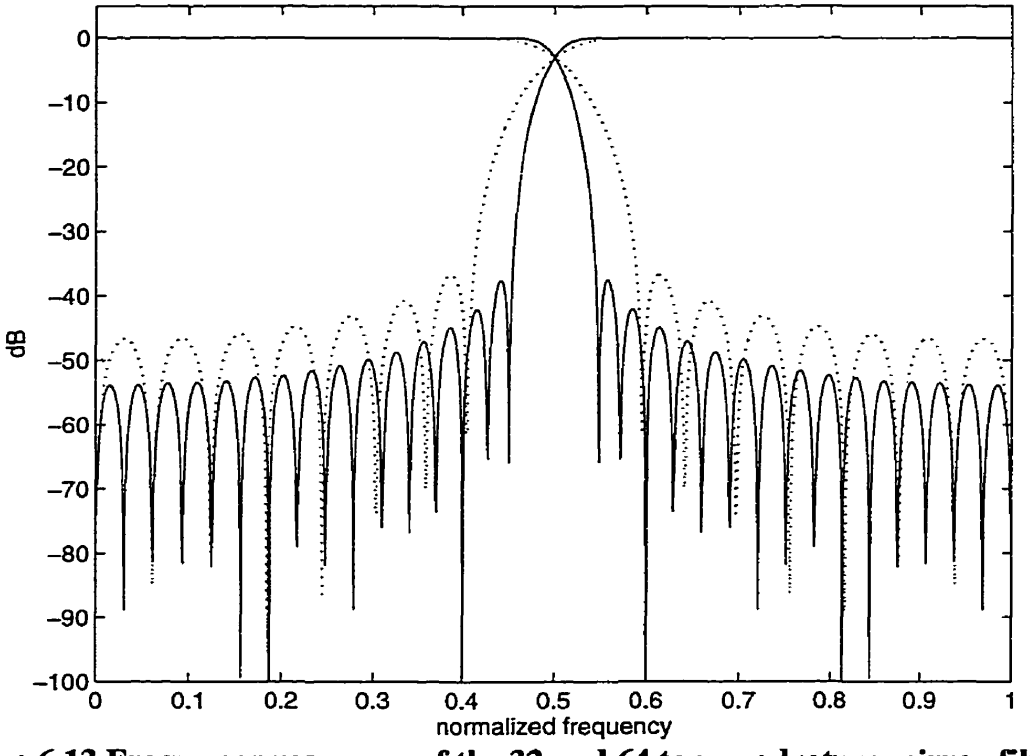


Figure 6.13 Frequency responses of the 32 and 64 tap quadrature mirror filters.

Two of Johnston's filters [110] were implemented and tested. The 32 tap prototype filter had a transition bandwidth of 0.0215π, a minimum stopband attenuation of 38 dB, and an amplitude reconstruction error of 0.025 dB (dotted line in Figure 6.13). Informal listening tests revealed that aliasing was audible under typical noise reduction conditions. The aliasing was perceived as a form of distortion superimposed onto the signal. Further tests using a 64 tap prototype filter indicated that the aliasing was reduced to an acceptable level with this design. The 64 tap filter had a transition bandwidth of 0.0115π, a

minimum stopband attenuation of 40 dB, and an amplitude reconstruction error of 0.025 dB (solid line in Figure 6.13). While the 64 tap filter has greater computational requirements than the 32 tap filter, the increase is very modest in comparison to the requirements of the overall noise reduction process.

By itself, subband based spectral subtraction offers little benefit over traditional spectral subtraction. However, when combined with sub-framing, a significant improvement can be obtained in the camera noise reduction process.

6.3 Spectral Subtraction using Sub-Framing and Subband Filtering

In the two previous sections the concepts of sub-framed and subband filtered spectral subtraction were introduced. The benefit of these approaches is that they offer an additional degree of flexibility in the noise reduction process which can improve its performance. More specifically, they allow the spectral subtraction process to be better matched to the cyclical characteristics of the noise. In this section sub-framing and subband filtering are combined in such a way that the spectral subtraction process can be matched to the time-frequency distribution of the camera noise. As a result, the noise reduction can be directed so that the noisier parts of the observed signal receive more aggressive processing while those portions of the signal with less noise receive less processing. As such, the underlying philosophy behind the subband/sub-frame approach is to minimize the processing which is applied to the input signal. This helps to minimize the artifacts which occur as a result of the processing.

In the subband/sub-frame approach the noisy signal is first divided into m subbands using a non-uniform quadrature mirror analysis filter bank as described in the previous section. Each subband signal $v_i(k)$ i=0,1,...,m-1 is then processed using the appropriate number of sub-frames for that band as described in Section 6.1. In this way, a different number of sub-frames can be applied to each frequency subband so that a good match can be obtained between the time-frequency decomposition of the spectral subtraction process and the time-frequency distribution of the camera noise. Typically, fewer sub-frames would be used at lower frequencies and more would be used at higher frequencies to coincide with the spectrogram of the camera noise seen in Figure 3.8. Each of the subband/sub-frame or time-frequency *cells* $v_{ij}(k)$ i=0,1,...,m-1; j_i=0,1,...,g_i-1 is processed by a separate spectral subtraction process $\Psi_{ij}(k)$. It is important to note that, as part of the spectral subtraction process, each time-frequency cell in decomposed further using a high resolution DFT. This can be viewed as applying a multi-channel uniform analysis bank to

each cell. The enhanced subband signals $\hat{s}_i(k)$ are then recombined via a synthesis filter bank to produce the output signal $\hat{s}(k)$. The subband/sub-frame method is shown in Figure 6.14.

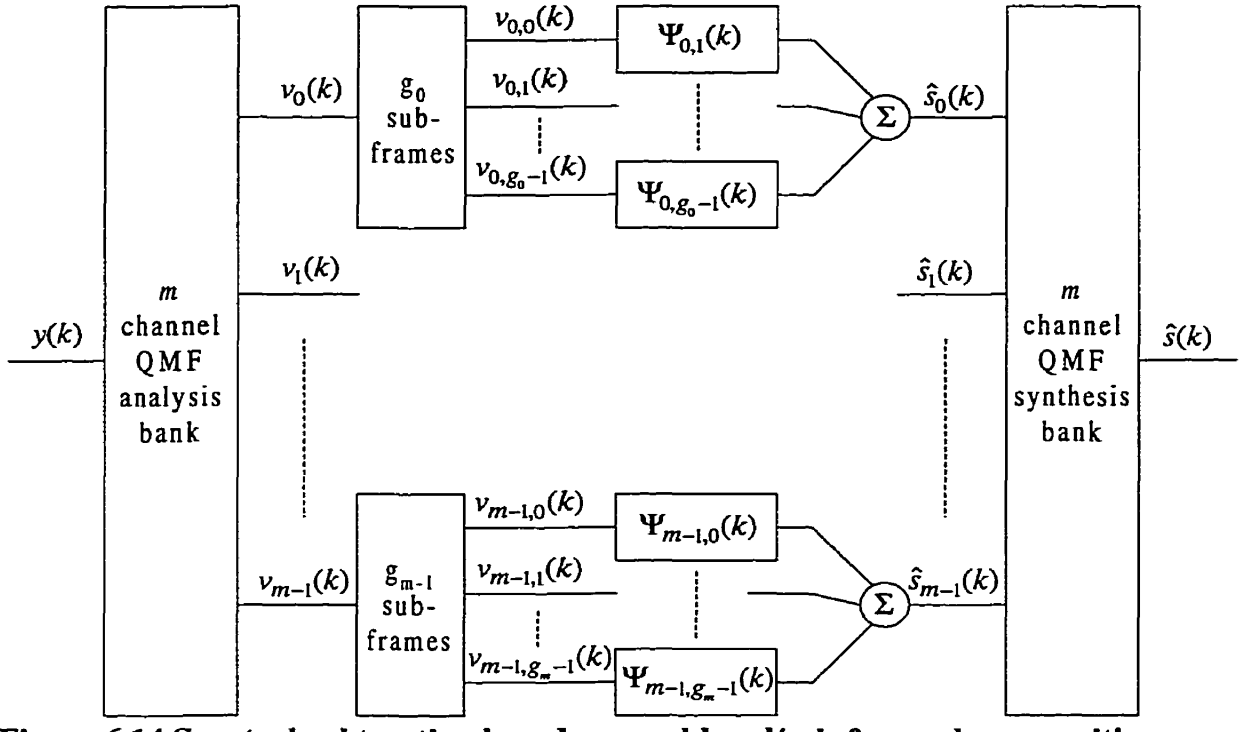


Figure 6.14 Spectral subtraction based on a subband/sub-frame decomposition.

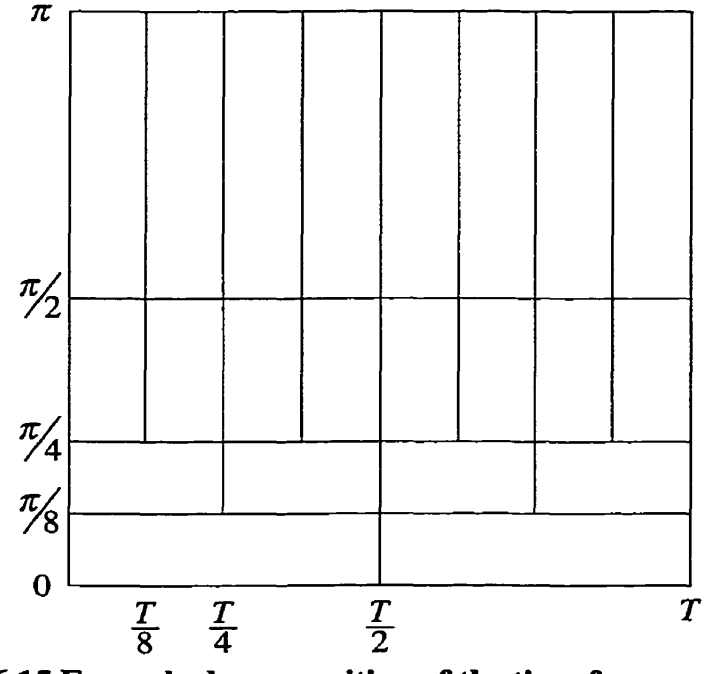


Figure 6.15 Example decomposition of the time-frequency plane.

Given the flexibility of the subband/sub-frame approach, the task is to determine the decomposition of the time-frequency plane which best matches the characteristics of the camera noise. One possibility which was explored is shown in Figure 6.15.

In this example, the input signal is decomposed into 4 octave-spaced subbands (i.e., m=4). The lowest frequency subband is then divided into 2 sub-frames (i.e., g_0=2). The next subband (from $\pi/8$ to $\pi/4$) is divided into 4 sub-frames (i.e., g_1=4). The remaining two subbands are divided into 8 sub-frames (i.e., g_2=g_3=8). This time-frequency decomposition provides a reasonable match to the characteristics of the camera noise.

To evaluate the suitability of a given time-frequency decomposition, it is instructive to examine the noise estimates $E[|N_{ij}(\omega_n)|]$ i=0,1,...,m-1; $j_i = 0,1,...,g_i - 1$ for each cell. These are shown in Figure 6.16 and Figure 6.17 for the above time-frequency decomposition.

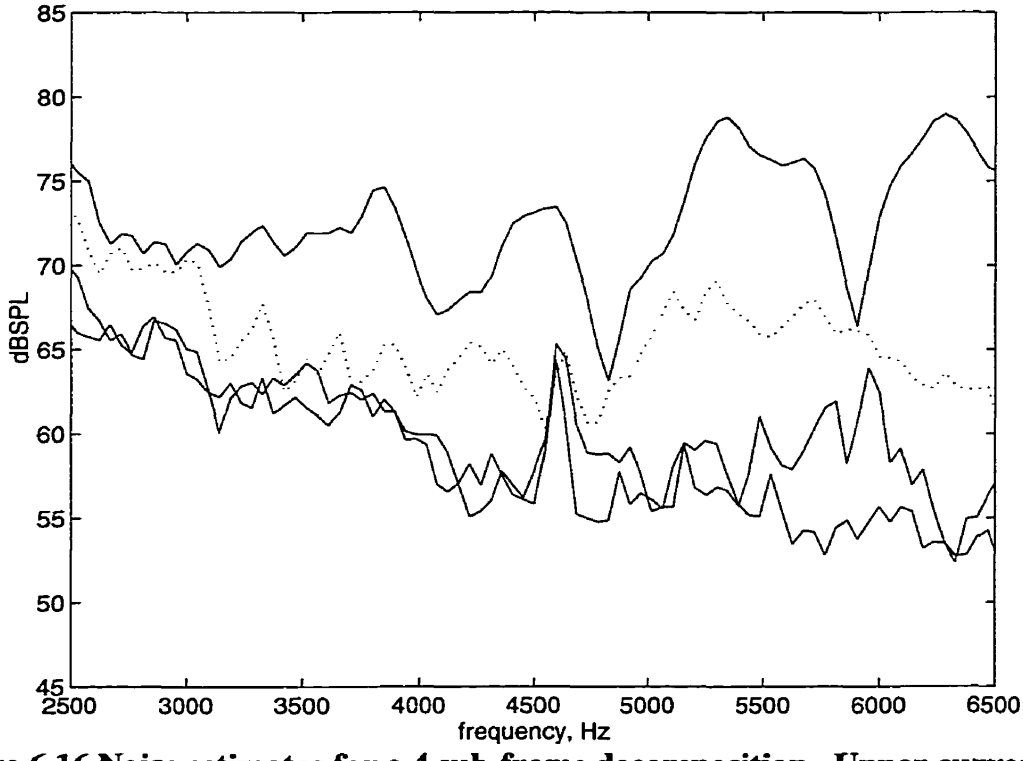


Figure 6.16 Noise estimates for a 4 sub-frame decomposition. Upper curve: peak sub-frame, middle curve: intermediate sub-frame, lower curves: null sub-frames.

Figure 6.16 shows the camera noise estimates for subband $v_1(k)$ $\pi/8$ to $\pi/4$ (or 3 to 6 kHz for a sampling rate of f_s=48kHz). The subband is divided into 4 sub-frames. The top curve in the figure is the noise estimate $E[|N_{1,0}(\omega_n)|]$ for the sub-frame centered on the peak for the noise pulse. The dotted curve is the noise estimate $E[|N_{1,1}(\omega_n)|]$ for the fol-

lowing sub-frame, and the bottom two curves are the noise estimates $E[|N_{1,2}(\omega_n)|]$ and $E[|N_{1,3}(\omega_n)|]$ for the two remaining sub-frames. Across this frequency range, the difference in the level of the noise in the peak sub-frame and the following sub-frame is about 7 dB. However, the noise estimates for the two remaining sub-frames are about 10 to 15 dB lower than the peak sub-frame. Therefore, the spectral subtraction processes $\Psi_{1,2}(k)$ and $\Psi_{1,3}(k)$ working on these sub-frames operate at a signal-to-noise ratio which is 10 to 15 dB higher than during the peak sub-frame. As a result, the artifacts which can occur due to the spectral subtraction process will be dramatically reduced for these sub-frames.

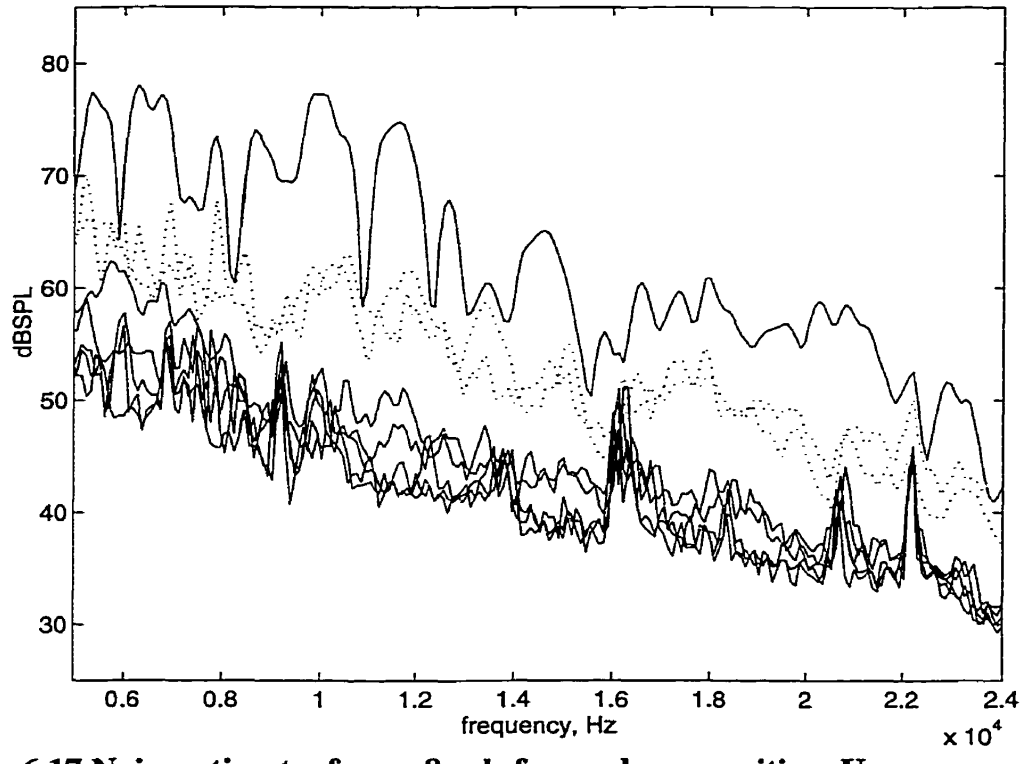


Figure 6.17 Noise estimates for an 8 sub-frame decomposition. Upper curve: peak sub-frame, dotted curves: intermediate sub-frame, lower curves: null sub-frames.

Figure 6.17 shows the noise estimates for the two subbands $v_2(k)$ and $v_3(k)$ ranging from $\pi/4$ to $\pi/2$ and $\pi/2$ to π (or 6 to 12 kHz and 12 to 24 kHz for a sampling rate of f_s=48kHz). These subbands are divided into 8 sub-frames. The top curve represents the noise estimate for the sub-frame which is centered on the peak of the noise pulse. The two dotted curves represent the noise estimates for the sub-frames which immediately follow the peak. The five lower curves are the noise estimates for the 5 remaining sub-frames. It can be seen that, for these two subbands, the noise estimate corresponding to the peak of the noise pulse (top curve) is about 10 dB higher than for the two following sub-frames, and about 20 dB higher than for the 5 remaining sub-frames. Therefore,

more aggressive processing can be concentrated on the peak sub-frame, and for the 7 remaining sub-frames, a dramatic reduction (10 to 20 dB) in the amount of processing applied to the signal can be achieved. As a result, the level of the artifacts produced by the spectral subtraction process is significantly reduced in these frequency sub-frames.

In the decomposition of the time-frequency plane shown in Figure 6.15, all of the sub-frames in a given subband were the same size. However, a subband can be divided using non-uniform sub-frames. In the noise estimates shown in Figure 6.16 (4 sub-frame case) the two lower curves are very similar and so there is no benefit in having separate sub-frames for this portion of the camera noise pulse. Therefore, it is sensible to combine these two sub-frames to form a single (larger) sub-frame. The resulting sub-frame has now doubled in length thereby doubling the frequency resolution of the noise estimate in this interval. Similarly, the noise estimates shown in the five lower curves of Figure 6.17 are very similar and so it is sensible to merge these sub-frames into a single sub-frame.

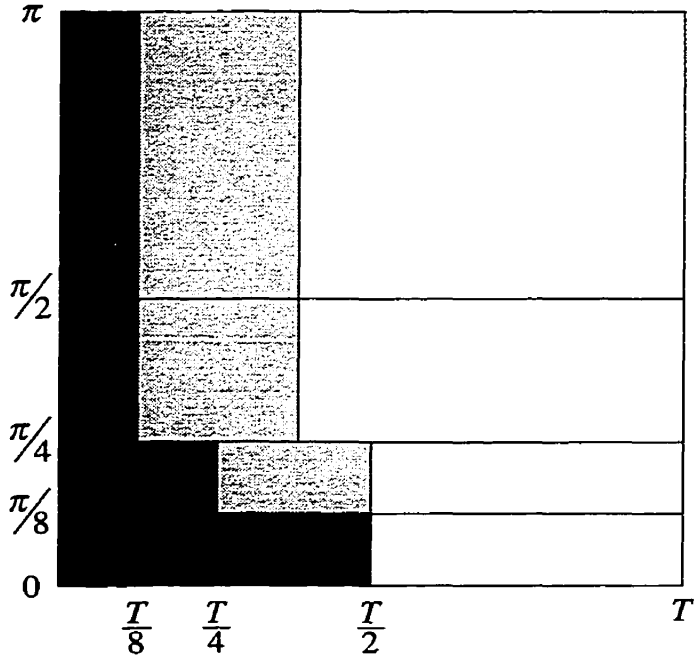


Figure 6.18 Decomposition of the time-frequency plane using non-uniform sub-frames.

This new decomposition of the time-frequency plane using non-uniform sub-frames is shown pictorially in Figure 6.18. The dark (black) portion of the figure represents those time-frequency cells which correspond to the peak portion of the camera noise pulse. The lighter gray portion represents the cells operating on the intermediate portion of the camera noise pulse (i.e., the dotted curves in Figure 6.16 and Figure 6.17). The white portion of the figure corresponds to the cells operating on the part of the input signal where the

level of the camera noise is the lowest (i.e., the lower curves in Figure 6.16 and Figure 6.17).

Given the use of non-uniform sub-frames, some consideration must be given to the choice of windows used prior to performing the DFT for the spectral subtraction process. As seen earlier in equation (5.16), in order to reconstruct an estimate of the clean signal $s(k)$ we require that the sum of the windowing functions be equal to 1. Recall that (5.16) implies the use of a rectangular synthesis window. When windowing sub-frames of different lengths, care must be taken to ensure that this basic requirement is met.

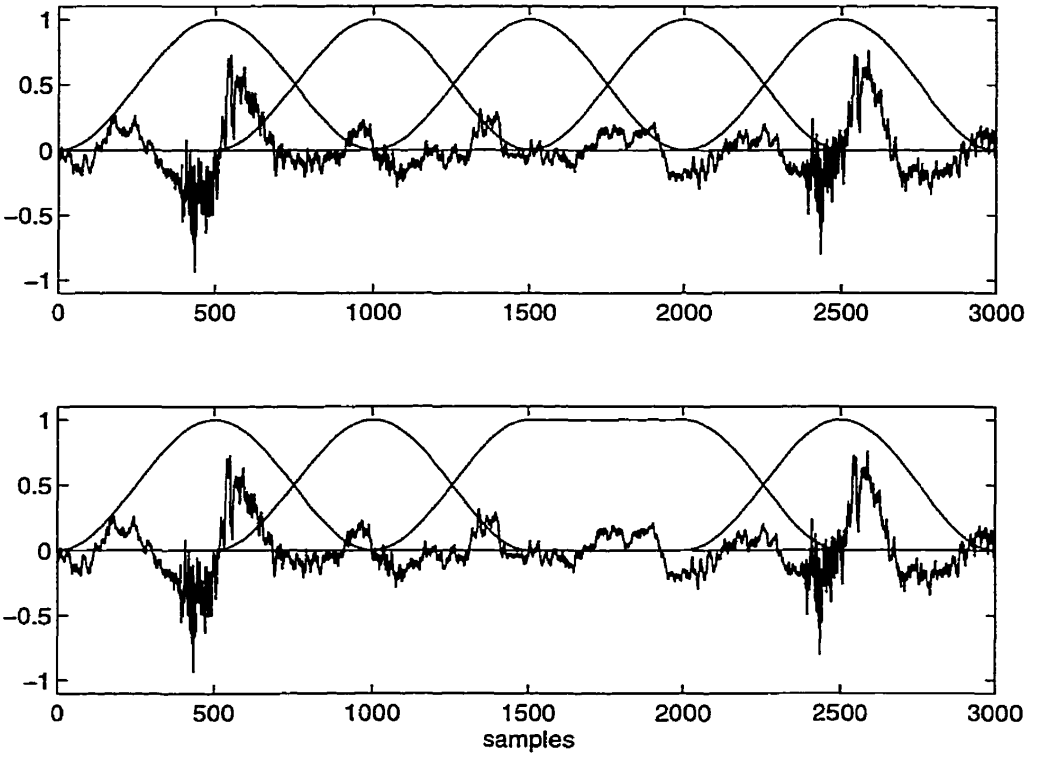


Figure 6.19 Uniform and non-uniform windowing based on 4 sub-frames.

The most straightforward approach for dealing with this problem is to use a form of hybrid window as shown in the lower plot of Figure 6.19. The upper plot shows one period of the camera noise divided into 4 sub-frames of equal length. The first window (sub-frame) is centered on the peak of the noise pulse, and is followed by three more windows before the next peak arrives. The windows are overlapped by 50%. In the lower plot, the third and fourth windows (sub-frames) are combined using a hybrid window. The rising slope on the left-hand side of the window is the first half of a Hanning window. The falling slope on the right-hand side of the window is the second half of the Hanning window. Between the two halves of the Hanning window is a flat region.

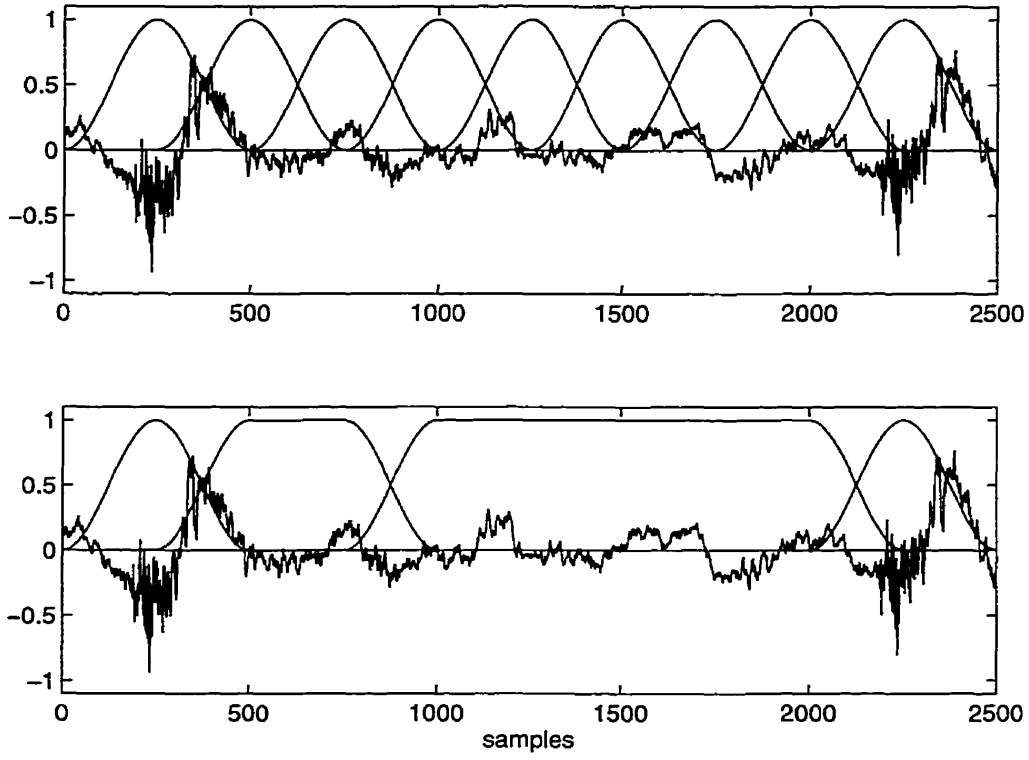


Figure 6.20 Uniform and non-uniform windowing based on 8 sub-frames.

Figure 6.20 shows a similar non-uniform windowing strategy for an 8 sub-frame decomposition. The upper plot shows the camera noise decomposed using 8 equal length windows. Again the windows overlap by 50%. In the lower plot, two different hybrid windows can be seen. The two windows differ in the length of the flat region of the window.

The hybrid windows can be described mathematically as follows,

$$w_1(k)=\begin{cases}0.5-0.5\cos(\pi k/K_1), & 0\le k\le K_1-1\\ 1, & K_1\le k\le 2K_1-1\\ 0.5-0.5\cos(\pi k/K_1), & 2K_1\le k\le 3K_1-1\end{cases} \tag{6.54}$$

where $K_1=T/4$ and

$$w_2(k)=\begin{cases}0.5-0.5\cos(\pi k/K_2), & 0\le k\le K_2-1\\ 1, & K_2\le k\le 5K_2-1\\ 0.5-0.5\cos(\pi k/K_2), & 5K_2\le k\le 6K_2-1\end{cases} \tag{6.55}$$

where $K_2=T/8$. Of course, window types other than the Hanning may be used in conjunction with the rectangular window. While the hybrid windows have the desired time characteristics (i.e., sum of the overlaps is equal to 1), they may not have appropriate fre-

quency characteristics when a perceptual model of the human auditory system is included in the processing. This matter will be considered in greater detail in the next chapter.

The underlying philosophy behind the subband/sub-frame spectral subtraction approach is to match the time-frequency decomposition of the signal to the characteristics of the noise. This allows the overall amount of processing which is applied to the input signal to be minimized. This, in turn, tends to minimize the artifacts which occur as a result of the noise reduction process. This philosophy will be extended further in the next chapter where a model of the ear is used to direct the noise reduction process such that the minimum processing necessary from a perceptual point of view is applied.

6.4 Interpretation of Subbands and Sub-Frames in Terms of Wavelets

In the preceding section it was seen that the performance of the spectral subtraction based noise reduction could be improved by matching the processing to the time-frequency distribution of the noise. This was accomplished by filtering the input signal using a non-uniform QMF analysis bank and then sub-framing the resulting subband signals. This time-frequency decomposition of the signal shares many similarities to a decomposition using wavelets or wavelet packets.

In this section, some fundamental aspects of wavelets will be described and the similarities to the time-frequency decomposition outlined in the previous section will be highlighted. As there are numerous articles and textbooks devoted to the topic of wavelets, we shall not endeavor to provide a thorough treatment of the topic. Rather, the discussion will be limited to a level wherein the reader can appreciate the relation between wavelets and the time-frequency decomposition developed earlier in this chapter. Comprehensive discussions on the topic of wavelets can be found in [113,114,115,116,111].

The scalar product of two signals $y(t)$ and $\varphi(t)$ in the $L_2(\mathbf{R})$ space of continuous-time energy functions is defined as,

$$< y(t), \varphi(t) > \; = \int_{-\infty}^{\infty} y(t)\, \varphi^*(t)\, dt \tag{6.56}$$

The scalar product allows a signal $y(t)$ to be mapped from its current domain to a transform domain defined by $\varphi(t)$. Setting $\varphi(t)=\exp(j\omega t)$, results in the well known Fourier transform,

$$< y(t), e^{j\omega t} > = \int_{-\infty}^{\infty} y(t) e^{-j\omega t} dt \ . \tag{6.57}$$

In order to allow the transform to have some form of time dependency, a windowing function $w(t-\tau)$ can be added thus giving the short-time Fourier transform (STFT),

$$< y(t), e^{j\omega t} > = \int_{-\infty}^{\infty} y(t) w(t-\tau) e^{-j\omega t} dt \ . \tag{6.58}$$

The STFT maps the input signal $y(t)$ onto the time-frequency plane in a uniform manner. That is, the time-frequency resolution is fixed over the entire time-frequency plane. However, non-uniform mappings (see Figure 6.15) may be desired in order to obtain a multi-resolution analysis of the signal. This is provided for directly by the wavelet transform which is defined as,

$$< y(t), a, b > = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} y(t) \psi^*(\frac{t-b}{a}) \ dt \ , \tag{6.59}$$

where the parameter $\psi(t)$ is called the *mother wavelet* which has a bandpass characteristic, b is the *time shift* parameter, and a is the *scaling* parameter. The parameter a provides a trade-off between resolution in time and resolution in frequency.

Another way to view these transforms is in terms of filter banks. It is well known that the STFT can be viewed as a filter bank having uniformly spaced filters [111]. That is, as can be seen in Figure 6.21, they are constant bandwidth filters.

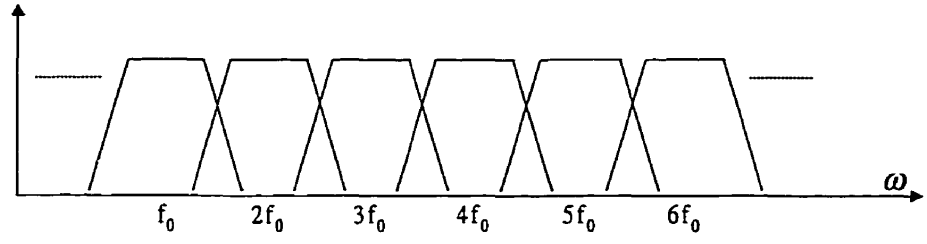


Figure 6.21 Uniformly spaced filters of the STFT.

The filters of the STFT filter bank are obtained by modulating (frequency shifting) a prototype lowpass filter $H_0(e^{j\omega})$,

$$H_n(e^{j\omega}) = H_0(e^{j(\omega-(2\pi n/N))}) \ . \tag{6.60}$$

The filters $H_n(e^{j\omega})$ n=0,1,...,N-1 are a series of bandpass filters. Flexibility in the STFT is obtained in the design of the prototype filter (basis function) which is determined by the choice of the window function.

The discrete wavelet transform (DWT) can also be viewed as a filter bank. However, the filters of the DWT are non-uniformly spaced. More specifically, they are constant Q filters,

$$Q=\frac{centre\ frequency}{bandwidth}\ .$$

The DWT filter bank is shown in Figure 6.22. This type of non-uniform filter bank was seen earlier in this chapter in the discussion of QMF banks (see Figure 6.11). Recall that for the QMF bank, the filtering was achieved using a binary tree structure based on iterations of a highpass and a lowpass filter.

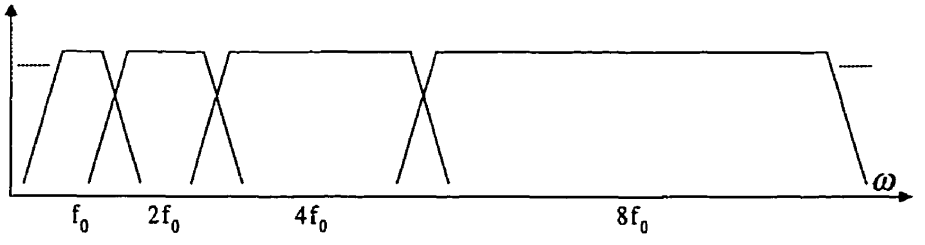


Figure 6.22 Non-uniform spaced filters of the DWT.

The filters of the DWT filter bank are obtained by first performing a frequency scaling of a prototype highpass filter $H(e^{j\omega})$,

$$H_n(e^{j\omega})=H(e^{ja^n\omega})\ . \tag{6.61}$$

The resulting filters are multiband rather than bandpass filters as was found with the STFT. In order to obtain a bandpass structure, the filters $H_n(e^{j\omega})$ are cascaded with appropriate lowpass filters $G_n(e^{j\omega})$. The resulting filter response for the nth filter bank channel is then

$$H(e^{ja^n\omega})G(e^{j\omega})\ \ n=0,1,\ldots,m\text{-}1\ , \tag{6.62}$$

where m is the number of channels in the filter bank.

This filter bank response can be obtained using the binary tree structure shown in Figure 6.23. Therefore, the structure of the DWT filter bank is the same as the QMF bank seen in Figure 6.11, and so the non-uniform subband decomposition described earlier can be viewed as a form of wavelet decomposition. The wavelet coefficients $u_l(k)$ $l=0,1,\ldots,m\text{-}1$ can be viewed as the subband signals $v_l(k)$ at the output of the non-uniform QMF analysis bank.

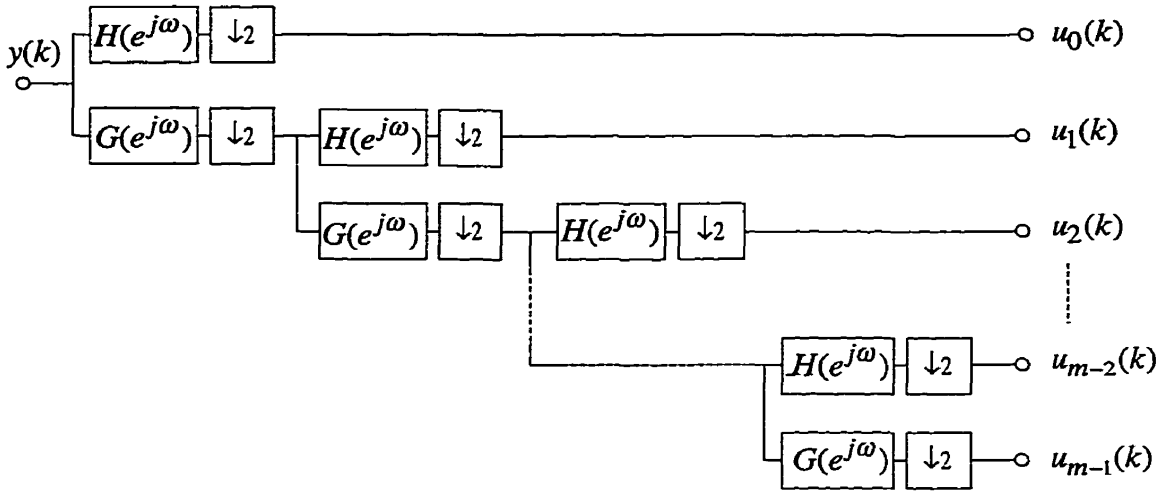


Figure 6.23 Filter bank representation of the DWT.

A generalization of the DWT is the concept of wavelet packets. First proposed by Coifman *et al.* [117,116], wavelet packets correspond to arbitrary tree-structured filter banks, and thus they allow the input signal to be decomposed in a far more flexible manner. The sub-framing and high resolution DFT which is performed on each of the subband signals as described in Section 6.1 could be viewed in terms of wavelet packets.

One of the merits of the wavelet transform is that it can be useful in unifying seemingly unrelated areas of research, and may therefore help to provide new insights [115]. While wavelets offer an interesting mathematical viewpoint for examining the time-frequency decomposition described in the previous section, in this thesis we prefer to use of the language of quadrature mirror filter banks and STFT's since they are more familiar to the electrical engineer. Moreover, the psychoacoustic models of the human auditory system which will be examined in the following chapter, are better understood in terms of STFT analysis. Also, perceptual-based processing of audio signals (e.g. compression) using wavelets has not tended to yield any improvement in performance over traditional STFT approaches [118,119].

6.5 Summary

In this chapter, the concepts of spectral subtraction using sub-frames and subband filtering were introduced. When steps are taken to synchronize the process to the film rate and to align the windows to the noise pulses, the combination of sub-frames and frequency subbands provides a significant improvement in the performance of the spectral subtraction algorithms. The benefit of these approaches is that they allow the spectral subtraction process to be better matched to the cyclical characteristics of the camera

noise. This inherently reduces all forms of audible artifacts since the average amount of processing applied to the signal is minimized. At the time intervals and frequencies where the camera noise is loudest, more aggressive processing is employed. Elsewhere, less aggressive noise suppression is applied. By using a combination of sub-frames and frequency subbands spectral subtraction can be used to successfully remove camera noise even for relatively poor signal-to-noise ratios.

7. SPECTRAL SUBTRACTION BASED ON MASKING IN THE HUMAN AUDITORY SYSTEM

So far, the focus has been placed on finding ways of obtaining the best estimate of the camera noise which is corrupting a desired (speech) signal. This estimate of the noise is subtracted from the noisy signal to give an estimate of the desired signal. While this is a sensible mathematical approach, Tsoukalas *et al.* [120] describe a novel modification to the process which can significantly improve the performance of a spectral subtraction based noise reduction system. They note that, due to the *masking* properties of the ear, it is not necessary to remove the entire noise component in a noisy signal. Rather, it is only necessary to remove that part of the noise which is perceptually relevant.

By limiting the noise reduction process to removing only the audible portion of the noise, the overall amount of noise reduction can be significantly reduced. This in turn, reduces the audible artifacts (musical noise in particular) which can result from the spectral subtraction process. This approach is in concert with the philosophy of processing in subbands and sub-frames in order to concentrate the processing on those parts of the signal which require it most, thus reducing the overall amount of processing applied to the signal.

It should be noted that Tsoukalas *et al.* were not the first to suggest that adding a perceptual model could improve the performance of a speech enhancement system. In 1981, Peterson and Boll [121] described a "perceptual subtraction" algorithm, which is essentially a spectral subtraction algorithm operating in the perceptual domain rather than the Fourier domain. Peterson and Boll state that their perceptual subtraction algorithm eliminates the musical noise. Interestingly, this work appears to have been largely ignored by other researchers. Cheng and O'Shaughnessy [122] also took advantage of the masking properties of the ear in their speech enhancement algorithm. We will focus our attention on the work of Tsoukalas *et al.* since it is the most recent and the most comprehensive.

In this chapter we review the basic features of the spectral subtraction method proposed by Tsoukalas *et al.* which is based on the perceptual audio quality measure (PAQM) developed by Beerends and Stemerdink [123]. This approach, which uses the critical band approximation to the human auditory system, has certain fundamental limitations which will be outlined. Due to these limitations, we choose to examine another

auditory model which is perhaps better suited to the task of noise reduction. Whereas PAQM combines many aspects of the peripheral auditory system into a computationally efficient model, the new method uses a more direct approximation of the various components of masking.

7.1 An Introduction to Auditory Masking

Perceptual based spectral subtraction relies on the masking properties of the ear. Stated simply, masking is the process by which one signal which would otherwise be audible is rendered inaudible by the presence of another signal (masker). Masking is a psychoacoustic phenomenon which occurs due to the nonlinearities in the human peripheral auditory system (i.e., the outer, middle, and inner ear).

Modern theories of masking derive directly from the work of Fletcher [124]. Fletcher measured the threshold at which a sinusoidal signal could be detected in the presence of bandpassed noise centered on the signal. He noted that the threshold of detection of the signal increased as the bandwidth of the noise was increased. However, he found that beyond a certain point any further increase in the bandwidth of the noise did not alter the signal detection threshold. To explain his findings, Fletcher reasoned that the peripheral auditory system must behave as a bank of overlapping bandpass filters which are now called the *auditory filters*. When detecting a sinusoidal signal in noise, the listener attends to (i.e., pays attention to) the auditory filter centered on the signal. Therefore, in Fletcher's experiments, if an increase in the bandwidth of the noise occurred within the bandwidth of the auditory filter, then the detection threshold increased. If however, the bandwidth of the noise was larger than the bandwidth of the auditory filter, the threshold of detection did not change. Fletcher called the bandwidth (of the noise) at which the threshold of detection no longer increases the *critical bandwidth*. Fletcher's model assumed that the auditory filters had perfect (i.e. rectangular) transfer functions. While this model is clearly unrealistic, Patterson and Moore [125] suggest that Fletcher realized that the details of the shape of the filters were of lesser importance than the general underlying concept.

Masking can be divided into two general classes: simultaneous and non-simultaneous. In simultaneous masking, the signal and the masker occur at the same instant in time, whereas in non-simultaneous masking the signal can arrive either before or after the masker [126]. Both forms of masking can be relatively difficult to measure and to quantify since they are highly nonlinear. For example, the auditory filters are asymmetric and

their widths vary across frequency. Moreover, the shape of the filters change significantly with the level of the input signal [126,127,128]. In the case of non-simultaneous masking, there is a nonlinear relation between the level of the masker and the amount of masking obtained. Also, due to phenomena such as beating; the amount of masking available is dependent on the spectral characteristics (noise-like versus tone-like) of the signal and the masker. Moreover, the masking effects of two or more maskers do not add linearly. Finally, masking thresholds can change dramatically depending on whether the listener is using one or both ears. How these various aspects of masking are dealt with is an important consideration in the accuracy of a given perceptual model.

7.2 Method of Tsoukalas et al

As mentioned earlier, Tsoukalas *et al.* added a psychoacoustic model to the spectral subtraction process. Their original work was based on the psychoacoustic model described by Johnston [129,], but more recently they have used the model found in PAQM [130]. The essence of the method can be summarized by the expression,

$$|\hat{S}_w(e^{j\omega})|^2 = |Y_w(e^{j\omega})|^2 - \max\{E[|N_w(e^{j\omega})|^2] - AMT,\ 0\}. \tag{7.1}$$

At each frequency the auditory masking threshold (*AMT*) due to the clean signal is calculated. The *AMT* for a given signal is the level below which all other signals will be masked. This threshold is then compared to the noise estimate for that frequency $E[|N_w(e^{j\omega})|]$. If the noise estimate falls below the *AMT*, then no processing is done at that frequency. If the noise estimate is above the *AMT*, then the difference between the noise estimate and the *AMT* is subtracted from the spectral magnitude of the noisy signal $|Y_w(e^{j\omega})|^2$. This reduces the noise level to the *AMT* thus bringing it to the threshold of audibility. Therefore, this version of spectral subtraction only removes the audible portion of the noise. Tsoukalas *et al.* claim that this results in significant improvements to the performance of the noise reduction process.[†]

It should be noted that Tsoukalas *et al.* do not calculate the *AMT* explicitly. Rather they calculate the compressed loudness function of the clean and noisy signals, and compare the two [128,123]. This approach stems from the use of the PAQM model which was designed as an objective means of evaluating the perceptual effects of noise in an audio signal. Therefore, the expression in (7.1) is only an idealized approximation to the

[†] Peterson and Boll [121] claim that including a perceptual model in their spectral subtraction process significantly reduced musical noise.

Tsoukalas method. Also, the above expression assumes that the clean signal $s(k)$ is available in order to determine the *AMT*. Of course, this assumption is not valid in practice, and so the power spectrum of the clean signal must be estimated by some means.

In order to determine the *AMT*, the signal $s(k)$ must be processed by some form of perceptual model. The choice of model is an important consideration since it determines the predicted *AMT* and thus the performance of the noise reduction system. The most common model, and indeed the one used by Tsoukalas *et al.*, is the one based largely on the work of Zwicker [127,128] and implemented in the context of a perceptual audio coder by Johnston [129]. This model, which is based on the concept of critical bands, has also been used by other researchers investigating the use of perceptual models in spectral subtraction [131,132]. Another perceptually based objective measurement system, NMR (noise masking ratio) is also based on the critical band model. In the present work, we choose to develop a new model based on the more recent psychoacoustic research conducted by Patterson and Moore [125,126] which provides a more direct approximation to the components of the peripheral auditory system. This approach has certain advantages in some applications.

7.2.1 The Critical Band Model

There are several steps in calculating the *AMT* using the critical band model. The signal $s(k)$ is first windowed (a Hanning window is typically used [129,130,133]) and transformed using an FFT of length N. The power spectrum $|S(e^{j\omega_n})|^2$ is then determined for that block. A critical band analysis is performed wherein the power spectrum is partitioned into critical bands according to the expression,

$$B_i = \sum_{n=bl_i}^{bh_i} |S(e^{j\omega_n})|^2 \ , \tag{7.2}$$

where bl_i is the lower boundary, bh_i is the upper boundary, and B_i is the energy of the ith critical band. Typically $i = 1,2,\ldots,25$ for full bandwidth signals and the boundaries of the critical bands are those defined by Scharf [134]. In general, the number of FFT bins will be much larger than the number of critical bands and so the critical band model reduces the frequency resolution. Johnston [129] points out that a true critical band analysis would sum the energy across one critical band for each frequency ω_n, n=0,1,...,N-1. Therefore, the 25 band implementation inherently discards some accuracy in its prediction of the masking threshold. It is important to note that the implementation of the criti-

cal band model described by Johnston was motivated by its application to perceptual audio coders and so it may not be optimal for the purpose of noise reduction.

The signal B_i is now in the bark domain, where 1 bark = 1 critical band. In order to account for the masking across critical bands, a *spreading function* is applied to the critical bands to produce the spread masking threshold C_i. The spreading operation consists of a convolution of B_i with an asymmetric triangular function H_{ij} having a lower slope of 25 dB/bark and an upper slope of -10 dB/bark (shown later in Figure 7.8) [128,135]

$$C_i = H_{ij} * B_i \ . \tag{7.3}$$

The spread masking threshold must next be related back to the bark domain in order to obtain the critical band masking threshold. Johnston [129] points out that this operation can not be done directly due to possible numerical instabilities and so it is approximated by a sub-optimal renormalization process. Finally, an approximation to the absolute threshold of hearing is applied as a lower limit of the masking threshold in each critical band. A full description of the details of the critical band model can be found in [129].

The critical band model provides a reasonable first approximation to the masking threshold created by a given signal, and it has been used with success in the development of perceptual audio codecs. The model is also quite computationally efficient and straightforward to implement. However, the model makes numerous assumptions and has certain fundamental shortcomings which can limit its performance.

The critical band model assumes that the auditory filters have perfect transfer functions (i.e., they have rectangular shapes). As discussed earlier, this assumption is obviously invalid, and in fact, the shapes of auditory filters are known to be triangular-like (on a dB scale) [136,137,125]. Also, the auditory filters are known to be asymmetrical and vary nonlinearly with the level of the signal. The shape of the auditory filters and their dependence on level is frequency dependent. As stated earlier, it is inaccurate to assume that there are only 25 critical bands and this assumption limits the frequency resolution of the model. Moreover, the critical bands used in the model are based on the results of Zwicker and Scharf [134,128,138], and assume that the width of the critical bands are constant for frequencies below 500 Hz. Moore and Glasberg [139] reviewed the results of several studies and demonstrated that the bandwidths of the auditory filters decrease for frequencies below 500 Hz. Therefore, the assumption regarding constant bandwidth critical bands below 500 Hz appears to be incorrect.

There are other factors which suggest that a psychoacoustic model with higher frequency resolution will provide better performance. As part of their mandate, Task Group 10-4 of the ITU (International Telecommunications Union) has been evaluating and developing a perceptual based objective measurement system. The group has adopted a model which significantly outperforms both the PAQM and NMR models [140] and uses a much higher frequency resolution. Furthermore, the results of a recent study by Soulodre *et al.* [141] examining the performance of six state-of-the-art perceptual audio codecs clearly demonstrate that the performance of the codecs is directly linked to the frequency resolution used in the codec. Without exception, the performance of the codec increased as the frequency resolution of the model increased.

In the critical band model, the masking of a signal at one frequency by a signal at another frequency is approximated by the spreading function. In reality, this masking is due to the overlapping of the auditory filters and is thus the amount of masking is both level and frequency dependent. Neither of these aspects is accounted for directly in the critical band model. The convolution which provides the spreading across frequency implicitly sums the masking on a power basis. Studies by Penner [142,143] (for the non-simultaneous case) and Lutfi [144,145] (for the simultaneous case) have shown that the addition of masking does not obey a simple linear law and can be either significantly greater or less than predicted by linear addition. This finding is supported by the results of a study by Green [146]. Humes and Jesteadt [147] proposed a model for the additivity of masking which accounted for both the simultaneous and non-simultaneous cases.

The critical band model described by Johnston does not account for non-simultaneous masking. This is an important omission and should be included in any model. It should be noted that Tsoukalas *et al.* do include a parameter to crudely account for one component of non-simultaneous masking (forward masking). Modeling the masking properties of the peripheral auditory system in terms of non-overlapping critical bands and a fixed spreading function is a somewhat indirect approach. As a result, it can be difficult to separate the contributions of the individual components of the auditory system, and therefore it is difficult to modify a given component independently of the others. As such, modifying the critical band model to account for the deficiencies listed above is not a trivial task. Moreover, it may be desirable, in some applications, to model the peripheral auditory system of a given individual (for example, in a noise reduction system as part of a hearing aid). This would be more difficult to achieve using the critical band model than by a more direct implementation. Finally, the psychoacoustic model described above, as

well as the model used by Tsoukalas *et al.* is only valid for monophonic signals. Due to binaural masking level differences (BMLD) large differences in masking can occur for stereo signals versus monophonic signals.

There are several parameters in the PAQM model (employed by Tsoukalas *et al.*) which are intended to represent cognitive rather than physiological aspects of the human auditory system. To calibrate these parameters within PAQM, Beerends and Stemerdink relied on the results of formal subjective evaluations of perceptual audio codecs. The calibration of these parameters therefore relies on the data set used in these tests and may not be valid for other data sets. That is, by tuning the psychoacoustic model to a given data set, the performance of the model may not translate well to other data sets. Also, the need to calibrate these parameters makes it very difficult and time consuming to make an individualized psychoacoustic model. Tsoukalas *et al.* used PAQM and NMR to tune the parameters of their model and therefore suffer from similar limitations.

7.3 Development of a New Psychoacoustic Model

In this section a new (high frequency resolution) psychoacoustic model is developed which is based largely on the work of Patterson and Moore and does not make some of the assumptions made in the critical band model. The new model implements the various components of the peripheral auditory model more directly and thus allows greater flexibility in modifying its characteristics.

Before deriving the model based on the work of Patterson and Moore, we consider an intermediate model. Like the critical band model described above, this intermediate model is based largely on the work of Zwicker. However, the model does not use the concept of critical bands. Rather, it retains the high frequency resolution resulting from the FFT. We will therefore refer to this psychoacoustic model as the *high resolution Zwicker model.* This model shares many similarities with another perceptually based objective measurement system, PERCEVAL developed by Paillard [148,149,133] which has been shown to work as well as PAQM. This model is useful in that it allows us to make direct comparisons between the work of Zwicker and that of Patterson and Moore.

7.3.1 The High Resolution Zwicker Model

The first stage in the model is to simulate the effects of the outer ear which is composed of the pinna and the auditory canal. The pinna modifies an incoming sound to some extent (particularly at high frequencies) and is important in our ability to localize sounds.

However, the effects of the pinna are ignored in this model since they vary with the direction of arrival of a sound. Due to its structure, the auditory canal tends to amplify frequencies in the range from about 1 kHz to 4 kHz [150]. The task of the middle ear is to provide an efficient transfer of acoustic energy between the outer ear and the inner ear. The three small bones of the middle ear (popularly known as the hammer, anvil, and stirrup), provide an acoustic impedance matching between the tympanic membrane and the oval window of the cochlea. The middle ear is most efficient at transferring sounds in the 500 to 4 kHz frequency range and is quite inefficient at higher frequencies. Terhardt *et al.* [151] proposed the following expression to account for the attenuation effects A of the auditory canal and the middle ear,

$$A = -6.5e^{[-0.6(f-3.3)^2]} + 0.001f^4 \text{ dB}, \tag{7.4}$$

where f is the frequency of the input signal in kHz. The first half of the expression provides a slight boost to frequencies between 1 and 5 kHz, while the second half gives a high frequency roll-off. The curve defined by (7.4) is plotted in Figure 7.1.

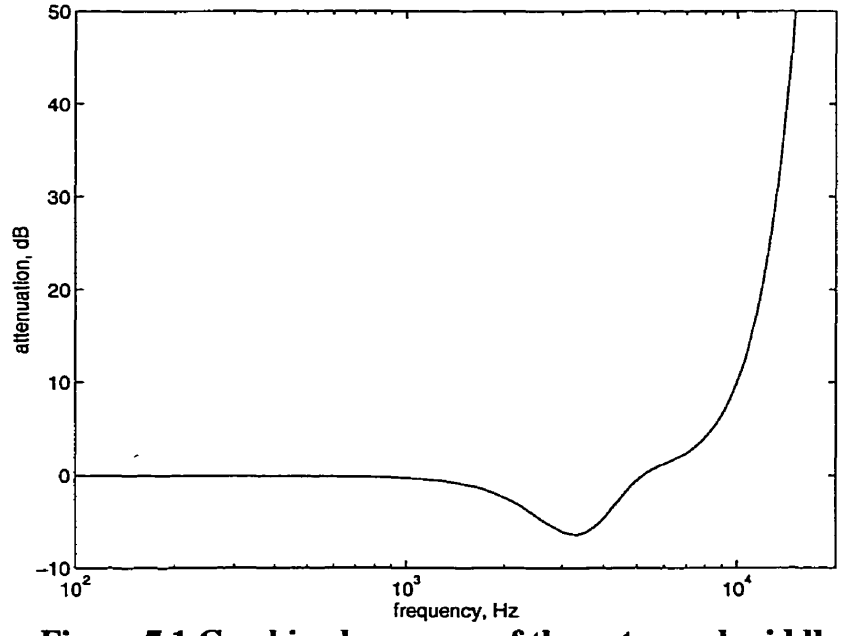


Figure 7.1 Combined response of the outer and middle ear.

A study by Shlien and Soulodre [107] showed that for many listeners the rate of high frequency roll-off given by (7.4) is too severe. They measured the threshold of hearing of 12 subjects at 500 Hz intervals in the frequency range from 4 kHz to 24 kHz. The results of these measurements for 5 subjects are shown in Figure 7.2 with the bold curve representing the expression proposed by Terhardt *et al.* for an average listener. It can be seen that equation (7.4) underestimates the high frequency acuity of some subjects and overes-

timates the acuity of others. Therefore, in applications where a customized perceptual model is warranted, user-specific measurements of the outer and middle ear responses will be necessary. Shlien and Soulodre found that the average of their data was in reasonable agreement with equation (7.4).

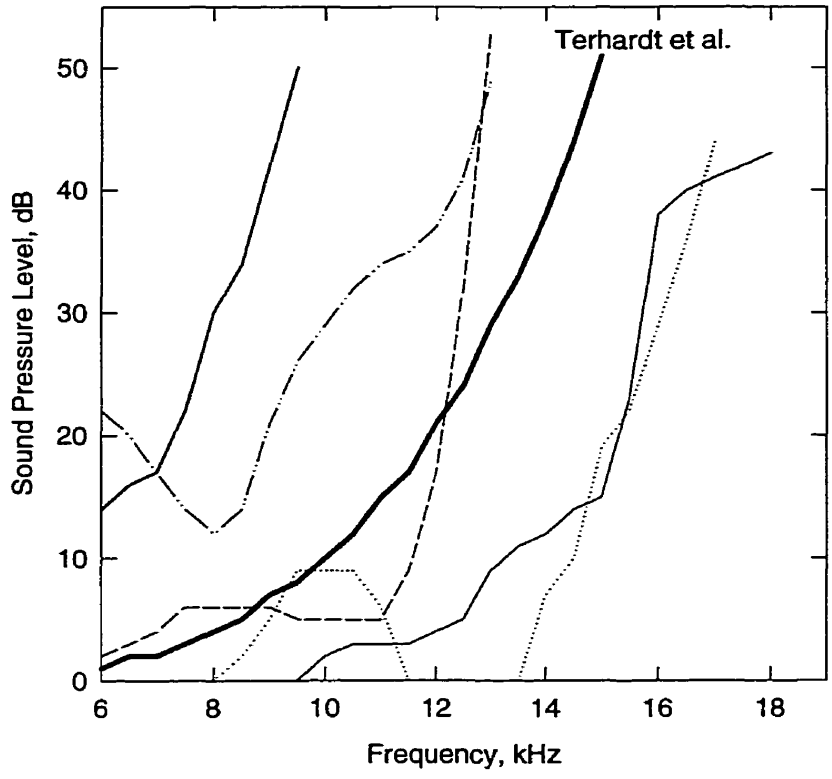


Figure 7.2 Absolute threshold of hearing for 5 subjects for frequencies above 6 kHz.

Another parameter of the peripheral auditory system which must be accounted for, is the internal noise of the inner ear. As in all real-world systems, there is an inherent noise floor in the auditory system. The noise is predominantly in the lower frequencies and Terhardt *et al.* [151] proposed the following expression to model it

$$noise\,floor = 3.64 f^{-0.8} \quad \text{dB}, \tag{7.5}$$

where f is the frequency in kHz. The inherent noise floor of the peripheral auditory system as proposed by Terhardt *et al.* is plotted as the solid curve in Figure 7.3. Included in the figure is the absolute threshold of hearing (dotted curve). It can be seen that, in this model, the absolute threshold of hearing is determined by a combination of the auditory canal response (outer ear), the middle ear response, and the internal noise of the inner ear. In the critical band model described by Johnston [129], the various components were accounted for at the last stage in the model where the final masking threshold was determined. This approach is inaccurate since the effects of the outer and middle ear should be

applied prior to the signal being mapped to the basilar membrane. Only the internal noise should be considered in the final stage of determining the masking threshold.

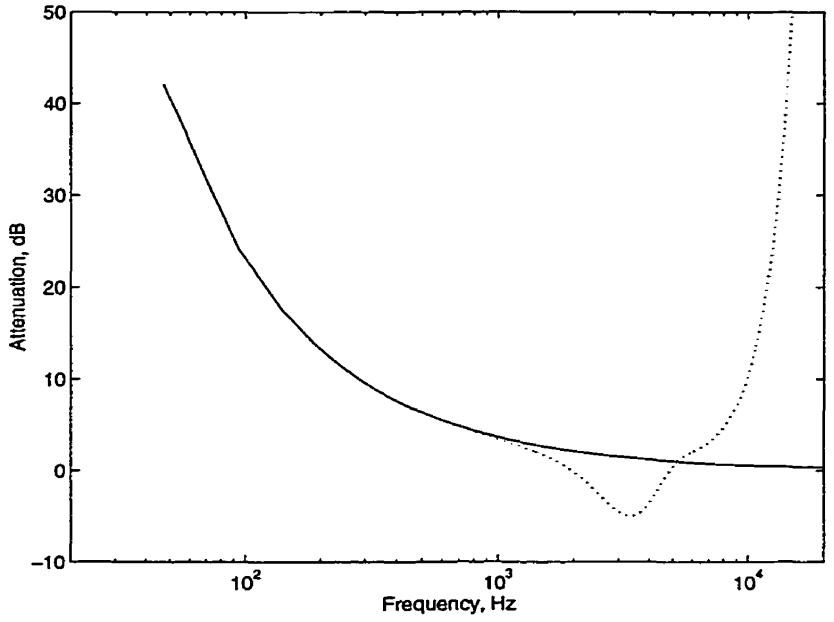


Figure 7.3 Internal noise of auditory system proposed by Terhardt *et al.*

There is evidence to suggest that the model proposed by Terhardt *et al.* is not entirely correct. That model assumes that there is no outer or middle ear filtering at frequencies below about 1 kHz. It assumes that the high threshold of hearing at low frequencies is due entirely to internal noise. However, Glasberg and Moore [152] point out that, at high sound pressure levels (100 phons), the equal loudness contours are not flat in the low frequencies. At these high levels, the signal is well above the noise floor of the auditory system and therefore some filtering of the signal seems to be taking place. This suggests that part of the threshold of hearing curve measured at frequencies below about 1 kHz is due to a filtering process in the middle ear, and part is due to the internal noise of the auditory system. To account for this, we derive a new model.

We begin by assuming that, at high sound pressure levels where the signal is well above the noise floor, the equal loudness contour represents the filtering action of the outer and middle ear. The resulting filter is based on the equal loudness contour (100 phons) provided in tabular form in ISO recommendation R.226 [152]. We thus propose the following analytic expression for the attenuating process A_s, representing the combined transfer functions of the outer and middle ear,

$$A_s = -6.5e^{[-0.6(f-3.3)^2]} + 0.001f^4 + 3.64f^{-0.8} - 80.64e^{[-4.712f^{0.5}]} \text{ dB}, \qquad (7.6)$$

where f is in kHz. This expression is plotted as the solid curve in Figure 7.4. The dotted curve in the figure represents the absolute threshold of hearing. It can be seen that for frequencies above 1 kHz, the absolute threshold of hearing curve is equal to the curve representing the filtering action A_s of the ear.

To derive an expression for the internal noise of the auditory system, we subtract the filter response A_s from the absolute threshold of hearing. Again, the tabular values provided in ISO recommendation R.226 were used in the calculations. In order to represent the internal noise of the auditory system, we propose the following expression,

$$N_{int}=80.64e^{-4.712f^{0.5}}\ \text{dB}, \tag{7.7}$$

where f is in kHz. This expression is plotted as the dashed curve in Figure 7.4.

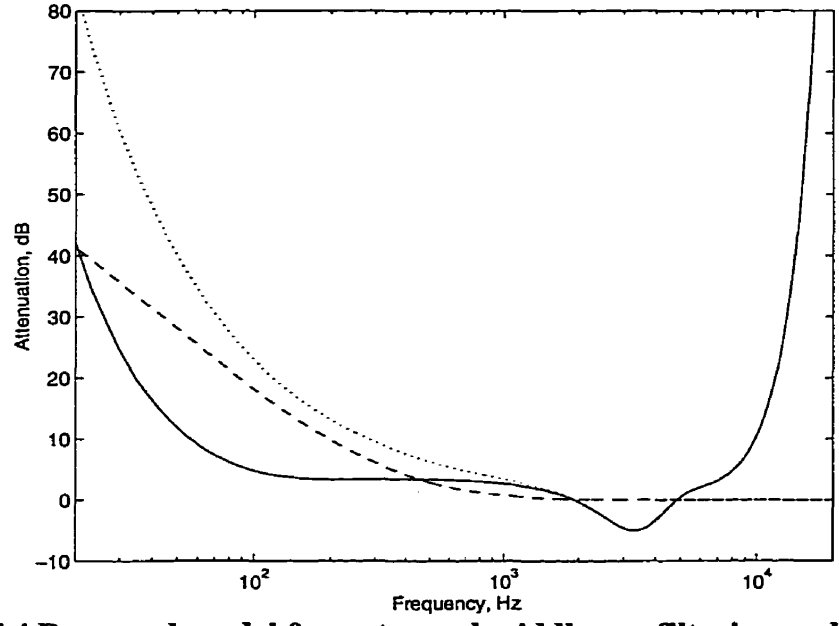


Figure 7.4 Proposed model for outer and middle ear filtering and internal noise floor. Solid curve: 100 phon equal loudness contour, dotted curve: threshold of hearing, dashed curve: internal noise floor.

The middle ear transfers the acoustic signal from the outer ear to the cochlea where the individual hair cells on the basilar membrane each respond to a particular frequency range. Effectively, the basilar membrane maps the linear frequency scale to a nonlinear pitch or mel scale. The mel is the unit of measure used to describe changes in pitch. Zwicker and Terhardt [153] proposed an analytical expression which approximates the mapping from the linear frequency scale to the bark scale. Since 1 bark equals approximately 100 mels, the expression can be easily modified to provide a mapping from frequency to mels.

$$mels = 1300\tan^{-1}(0.76f) + 350\tan^{-1}((f/7.5)^2)\,, \tag{7.8}$$

where f is the frequency in kHz. The curve defined by (7.8) is plotted in Figure 7.5.

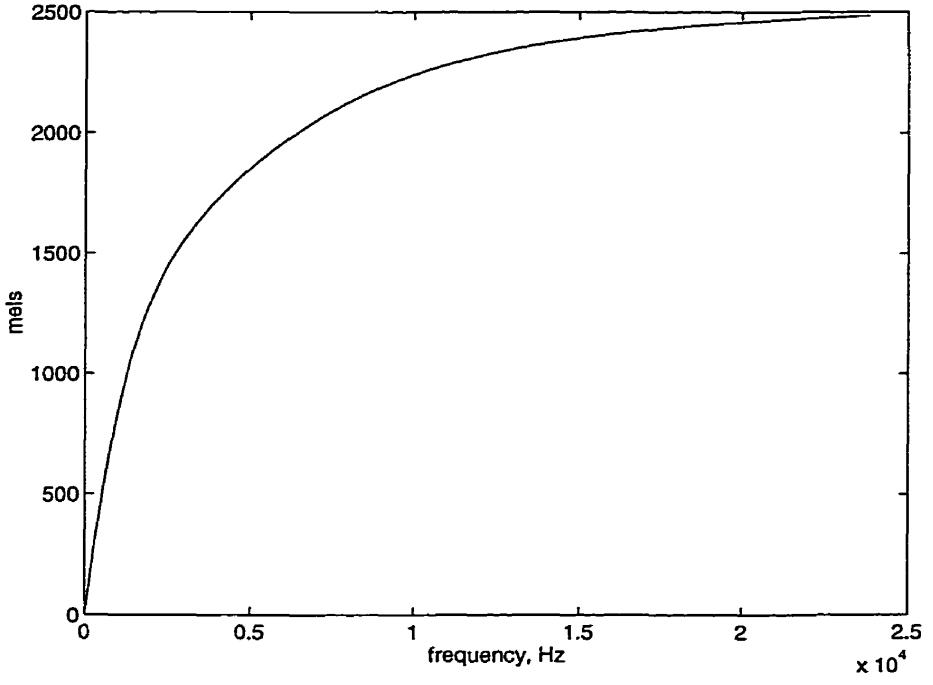


Figure 7.5 Mapping from linear frequency scale to mel (or bark) scale.

The curve suggests that the entire audible frequency range (about 20 Hz to 20 kHz) can be mapped into a range of 2500 mels which is equivalent to 25 bark. This observation serves as the basis for the earlier assumption (in the critical band model) that the frequency content of a signal can be divided into 25 critical bands. While the curve in Figure 7.5 is the basis of the division of the frequency scale into critical bands, there is some question as to its validity. Shlien and Soulodre [107] measured frequency to mel mapping functions of 15 subjects. This was done by measuring the subjects' ability to detect small variations in the frequency of a tone. Specifically, a frequency modulation threshold test as described by Zwicker and Fastl [128] was conducted. The test consisted of playing a tone whose frequency was modulated by $\pm\Delta f$ and asking the subject to indicate when he heard a wavering pitch, rather than a steady pitch. The threshold, Δf_T where the listener is able to detect a modulation in the pitch typically increases with increasing frequency. The results of these measurements for five of the subjects are plotted in Figure 7.6 in terms of the just noticeable variation in frequency (JNVF).

The JNVF, the critical bandwidth, and the frequency to distance mapping along the basilar membrane are all believed to be related [128,154]. Specifically, the JNVF is be-

lieved to map to a constant step size along the basilar membrane. The JNVF has also been shown to be related to the critical bandwidth by a constant factor of about 25.

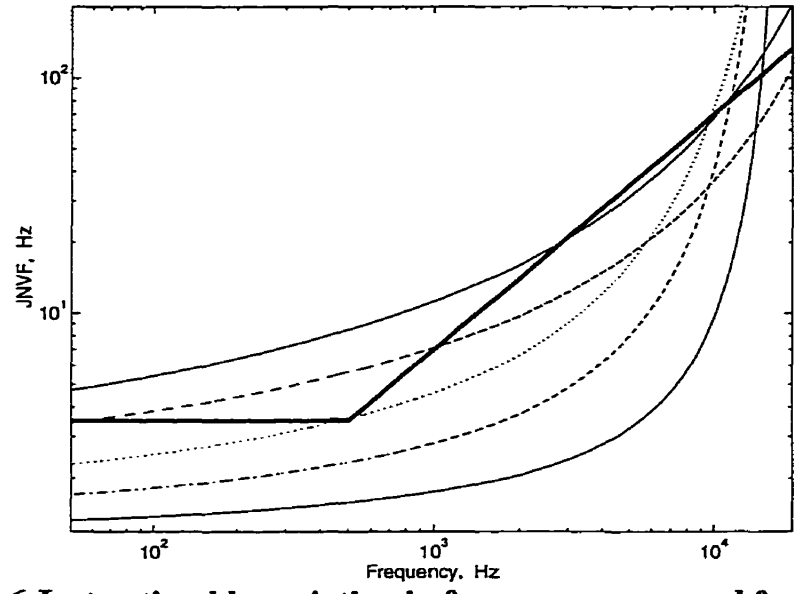


Figure 7.6 Just noticeable variation in frequency measured for 5 subjects.

From the results of the JNVF measurements, the frequency to mel mapping for each subject can be derived by integrating over the JNVF curve,

$$mel(f) = \int_0^f JNVF(f)\, df \ . \tag{7.9}$$

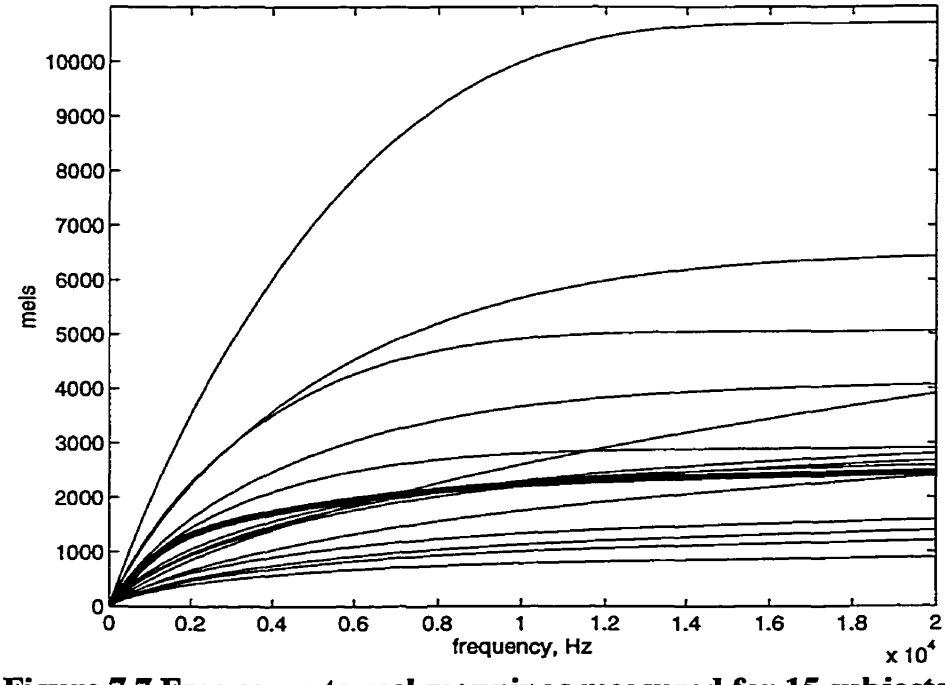


Figure 7.7 Frequency to mel mappings measured for 15 subjects.

The results of these measurements for all 15 subjects are shown in Figure 7.7. The bold line indicates the mapping proposed by Zwicker.

It can be seen from the figure that there are large deviations between the measured mappings and the curve proposed by Zwicker. Of particular interest are those curves which lie above the Zwicker curve. These curves represent subjects who demonstrated better frequency resolution than predicted by Zwicker and Terhardt. As such, for these subjects, the linear frequency scale maps onto a larger range of the mel scale. This suggests that these subjects require more than 25 critical bands in order to accurately describe their peripheral auditory systems. These findings are supported by the results of a study by Stevens and Volkmann [155] who found that, on average, the frequency range from 40 Hz to 12 kHz maps onto an interval of about 3500 mels, or about 35 bark. Another study which supports this finding is that of Moore and Glasberg [139,156] who found that the widths of the critical bands predicted by Zwicker are too large for frequencies below 500 Hz. This implies that subjects have better frequency resolution below 500 Hz than predicted by Zwicker. The derivation of a new perceptual model (different from the Zwicker model) is strongly motivated by the above findings.

In the critical band model, the simultaneous masking across frequencies is calculated using a spreading function. The spreading function consists of a two slope triangular function as shown in Figure 7.8. The spreading function is convolved with the signal mapped to the bark or mel domain.

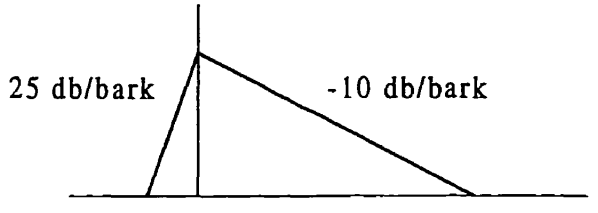


Figure 7.8 Spreading function proposed by Zwicker and Fastl.

The spreading function is typically assumed to have a rising slope S_1 of 25 dB/bark and a falling slope S_2 of about -10 dB/bark [135]. The rising slope represents the downward component of masking wherein a higher frequency signal masks a lower frequency signal. The falling slope represents the upward masking component wherein a lower frequency signal masks a higher frequency signal. Clearly, upward masking is the dominant effect. That is, lower frequency signals tend to better mask higher frequency signals rather than the converse. It is known that the amount of upward and downward masking varies with the level of the signal although the variation in downward masking with level

is not very large. To account for the variation in upward masking, the following expression is used by Beerends and Stemerdink [123] in the PAQM model for the falling slope of Figure 7.8,

$$S_2 = 22 + \min(\frac{230}{f},\ 10) - 0.2L \ \text{dB/bark} \tag{7.10}$$

where f is the masker frequency in hertz, and L is in dB SPL. PAQM uses a value of 31dB/bark for S_1. Terhardt *et al.* [151] propose a slightly different expression to account for the variation in the upper slope of the spreading function with level,

$$S_2 = 24 + \frac{230}{f} - 0.2L \ \text{dB/bark}. \tag{7.11}$$

As will be seen later, the choice of equation can significantly affect the amount of masking predicted by the model.

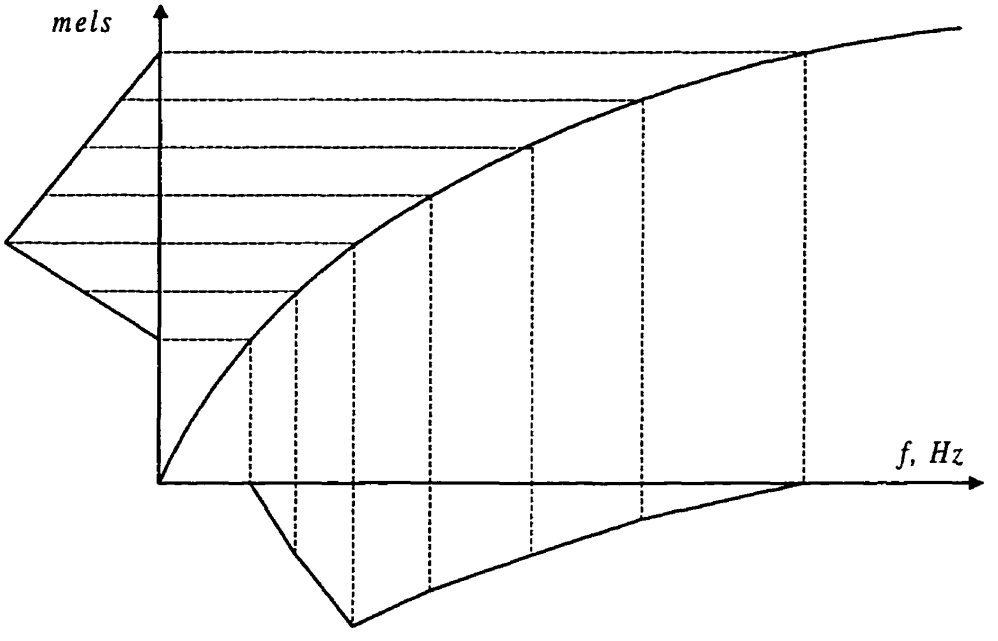


Figure 7.9 Mapping of spreading function in the mel domain to excitation pattern in the frequency domain.

Typically, in order to calculate the spread masking threshold, the spectral magnitude of the signal would be mapped to the bark domain using (7.8), and the spreading function would be applied. Rather than take this approach, we choose instead to perform all masking calculations in the linear frequency domain. This will allow for more direct comparisons between the Zwicker and the Patterson and Moore psychoacoustic models. To do this we map the spreading function at each point along the mel scale, to its equivalent location along the frequency scale. This is done using the frequency to mapping re-

lation defined in (7.8) in conjunction with the spreading function defined above. The process is shown pictorially in Figure 7.9.

The mapping depicted in Figure 7.9 provides an *excitation pattern* in the linear frequency domain. The excitation pattern refers to the level of excitation across frequency (or more specifically, across the basilar membrane) due to a given input signal and thus accounts for the effects of masking. The mapping process is determined for each position along the mel scale by sliding the spreading function along the mel scale and calculating the resulting excitation pattern. This yields a family of excitation patterns which can be summarized by the excitation matrix $\mathbf{E}_{zwicker}$, where each row represents the excitation pattern due to the frequency component of an input signal determined by the column of the matrix. Calculating the composite excitation pattern $\Im_s$ for an arbitrary input signal $s(k)$ is done through the following matrix multiplication,

$$\Im_s = \mathbf{s}_\omega \mathbf{E}_{zwicker} \tag{7.12}$$

where $\mathbf{s}_\omega$ is a 1 by N vector of the spectral magnitude coefficients of $s(k)$. $\Im_s$ is therefore a 1 by N vector containing a "smeared" version of $\mathbf{s}_\omega$ reflecting the effects of simultaneous masking.

The excitation matrix $\mathbf{E}_{zwicker}$, is shown graphically in Figure 7.10. Each curve in the figure represents the excitation pattern for a sinusoidal input signal. The front-most curve is the excitation pattern for the lowest frequency, while the furthest curve is for the highest frequency. The height of the curve represents the excitation level at each frequency.

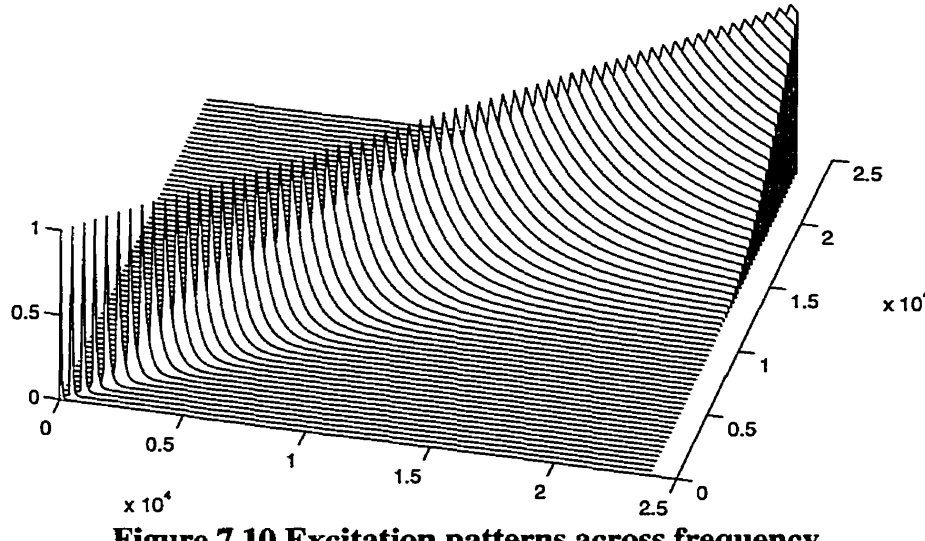


Figure 7.10 Excitation patterns across frequency.

In discussing Fletcher's work on critical bands, the concept of auditory filters was introduced. As one would expect, there is a direct link between the excitation pattern matrix $\mathbf{E}_{zwicker}$ and the auditory filters. In theory, the responses of the auditory filters are simply the columns of excitation matrix. That is,

$$\mathbf{AF}_{Zwicker} = (\mathbf{E}_{Zwicker})^T , \tag{7.13}$$

where $\mathbf{AF}_{zwicker}$ is a matrix whose rows are the auditory filters of the Zwicker model. It should be pointed out that, since the Zwicker model is not typically viewed in the linear frequency domain, the true auditory filters predicted by the model are not provided in the literature. Rather, the Zwicker model assumes the ideal filter characteristics associated with the concept of critical bands. Therefore, the present analysis provides an interesting new way of viewing the Zwicker model, and allows comparisons between this model and the Patterson and Moore model which is based on direct measurements of the auditory filters. This comparison will be done in the next section.

With the various components of the high resolution Zwicker model defined, it is now possible to calculate the simultaneous masking threshold for a given signal. As a test signal, a sum of 3 sinusoids at frequencies of 250, 1000, and 4000 Hz was used. This signal was chosen since Zwicker and Feldtkeller [127] provide measured masking thresholds for signals at these frequencies and so a direct comparison of the results is possible. The masking threshold for the test signal is plotted in Figure 7.11.

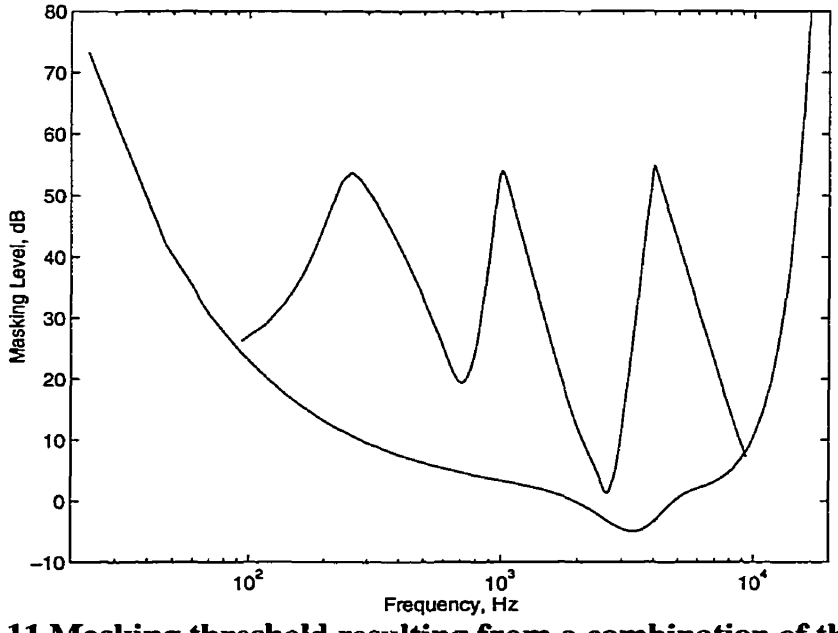


Figure 7.11 Masking threshold resulting from a combination of three sinusoids.

It can be seen from the figure that, as expected from the spreading function, the signal provides more masking for frequencies above the three centre frequencies than below. The lower limit of the masking threshold is the absolute threshold of hearing. The results of Figure 7.11 match very well with the measurements by Zwicker and Feldtkeller, thus suggesting that the model is providing a good prediction of measured masking thresholds.

7.3.2 The Patterson-Moore Model

In this section, we develop a new model based largely on the psychoacoustic studies of Patterson and Moore. The auditory filters and excitation patterns predicted by this model will be compared to the those predicted by the high resolution Zwicker model developed in the previous section.

As stated earlier, the human auditory system can be described as a filter bank composed of overlapping bandpass filters with the bandwidth of the filters increasing with increasing frequency. Masking occurs because of the overlapping of the filters which results in significant leakage of a signal into adjacent auditory filters. There are two basic approaches to measuring and predicting the masking properties of the human auditory system. One approach is to measure the amount of masking provided by a given masker signal [157,158,159]. This is referred to the masked audiogram approach. Here, a masker (typically a tone or a narrowband noise) is played and the threshold at which a probe signal (a tone) can be detected by a listener is measured across frequency. This method yields a prediction of the masking pattern due to the masker. To determine the composite masking pattern, the individual masking patterns for each frequency component of the masker are summed. While this method is intuitively appealing since it provides the desired masking information directly, it can yield erroneous results due to psychoacoustic phenomena such as beating and *off-frequency listening* [125]. The critical band model and the high resolution Zwicker model are both based on measurements using this approach for predicting the effects of masking.

The second approach is to measure the characteristics of the auditory filters at various frequencies across the audible range. Given the characteristics of the auditory filters, the excitation pattern for a signal is predicted by calculating the output of each auditory filter. Typically, the auditory filters are measured using the notched-noise method proposed by Patterson [136]. The notched-noise method has been shown to overcome the problems and beating and off-frequency listening and is considered to be a more reliable measure-

ment technique. The Patterson-Moore model is based on measurements of the auditory filters.

Patterson [160] measured the shapes of the auditory filters at 3 frequencies (0.5, 1.0, and 2.0 kHz) and showed that, to a first approximation, they were linear when plotted in decibels on a linear frequency scale. Based on this observation, Patterson *et al.* [137] suggested the following expression to describe the response of the auditory filters

$$W(g) = (1 + pg)e^{-pg} , \tag{7.14}$$

where g is the normalized distance from the centre frequency f_0 of the filter to the evaluation point,

$$g = |f - f_0| / f_0 . \tag{7.15}$$

The parameter p determines the rate of attenuation (i.e., slopes) of the filter and thus its bandwidth. A smaller value of p results in a slower rate of attenuation and thus a larger bandwidth filter. This parameter can be determined from the results of the notched-noise measurement method. Since the variable g is always positive, the resulting filter response appears as two back-to-back exponentials. The term $(1+pg)$ serves to round off the top of the curves where the two exponentials meet. As a result, the filter shapes suggested by Patterson *et al.* are referred to as the rounded exponential, or Roex filter shapes.

In the critical band model the auditory filters were assumed to be ideal rectangular filters, whereas the measurements by Patterson show that the filters are better approximated by a pair of back-to-back rounded exponentials. It is convenient in some instances to be able to express the shape and parameters of the Roex auditory filter in terms of an ideal rectangular filter. To do this, the bandwidth of the Roex filter can be expressed in terms of the equivalent rectangular bandwidth (ERB). The ERB of a given auditory filter is equal to the bandwidth of perfect rectangular filter which passes the same power of white noise as the auditory filter. That is, the area under the ERB curve is equal to the area under the Roex curve for a given auditory filter.

Numerous researchers have measured the auditory filter shapes for various frequencies. Moore and Glasberg [139] compiled the results from several studies and calculated the ERB's for each auditory filter measurement. They found very good agreement in the results across the studies and derived an analytic expression which successfully predicts the ERB's for auditory filters with centre frequencies between 125 Hz and 6.5 kHz

$$\text{ERB} = 6.23f^2 + 93.39f + 28.52 \text{ Hz} \quad 0.125 < f < 6.5 \text{ kHz} , \tag{7.16}$$

where f is the frequency in kilohertz. These results are valid for a notched-noise level of 40 dB SPL.

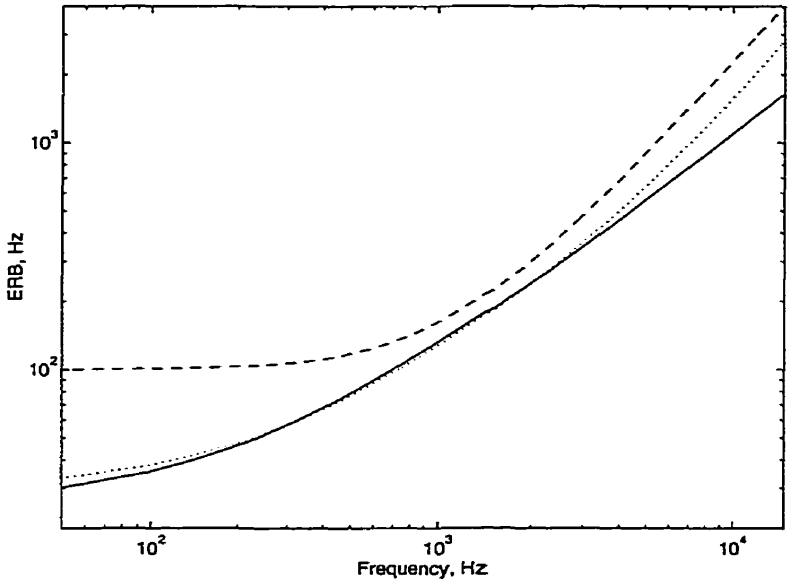


Figure 7.12 Equivalent rectangular bandwidth (ERB) as a function of frequency. Dotted curve: old ERB expression, solid curve: new ERB expression, dashed curve: bandwidth of critical bands.

More recently, Glasberg and Moore [152,161,162] have added the results from newer studies which allows the prediction of the ERB's to centre frequencies between 100 Hz and 15 kHz. The new expression for calculating ERB's is,

$$\text{ERB} = 24.7*(4.37*f+1) \text{ Hz} \quad 0.1 < f < 15 \text{ kHz} . \tag{7.17}$$

The two expressions for ERB are plotted in Figure 7.12. The dotted curve represents the older expression of (7.16), while the solid curve is the new expression (7.17). Also included in the figure (dashed curve) is the classical critical band function proposed by Zwicker and Terhardt [153].

It can be seen from the figure that the two ERB curves are very similar for frequencies below about 2.5 kHz. However, the two curves diverge with the newer expression (7.17) predicting smaller ERB's at higher frequencies. In applications which require a psychoacoustic model extending beyond 6 kHz, it is very important to use expression (7.17) rather than (7.15) since the older expression yields excitation patterns (and hence masking thresholds) which are significantly inaccurate at higher frequencies.

It can be seen that the critical band curve is quite different from the two ERB curves. In particular, the critical band estimates below 500 Hz are much larger than the ERB es-

timates. If the critical band model and the ERB model were essentially the same, then the curves would be parallel to each other. Since they are not, we expect the two models to produce significantly different predictions of masking, particularly at lower frequencies.

With the ERB at a given frequency as defined by (7.17), it is a straightforward process to determine the parameter p for the auditory filter and thus the shape of the auditory filter at each frequency can be determined. To do this, we recall that the area A_{ERB} under a given ERB filter curve is equal to the area A_{AF} under the corresponding auditory filter curve. A_{ERB} is simply

$$A_{ERB} = \frac{ERB}{f_0}, \tag{7.18}$$

where f_0 is the centre frequency of the auditory filter. A_{AF} is found by solving the following integral

$$A_{AF} = 2\int_0^\infty (1+pg)e^{-pg}\,dg \tag{7.19}$$

$$= 2\left|-p^{-1}(2+pg)e^{-pg}\right|_0^\infty \tag{7.20}$$

$$= \frac{4}{p} \tag{7.21}$$

Equating A_{ERB} and A_{AF} and solving for p, we get

$$p = \frac{4f_0}{ERB} \tag{7.22}$$

With equations (7.14), (7.17), and (7.22), it is a straightforward process to derive the shape of the auditory filter at any frequency.

It is of interest to compare the Patterson-Moore model to the Zwicker model described earlier to determine if they would predict the same masking threshold for a given input signal. Despite its fundamental importance, this type of direct comparison does not appear to be available in the literature. To compare the two models, it is convenient to derive a general expression for the shape of the Patterson-Moore auditory filters in terms of frequency. We begin with the following equations which were introduced earlier,

$$W(g) = (1+pg)e^{-pg}, \tag{7.23}$$

$$g = |f-f_0|/f_0, \tag{7.24}$$

$$\text{ERB} = 24.7\,(4.37f_0+1)\ \text{Hz} \quad 0.1 < f_0 < 15\ \text{kHz}, \tag{7.25}$$

$$p = \frac{4f_0}{ERB}. \tag{7.26}$$

Substituting (7.24) into (7.23) gives,

$$AF_{PM}(f,f_0)=(1+\frac{p|f-f_0|}{f_0})e^{(-\frac{p|f-f_0|}{f_0}g)}, \tag{7.27}$$

where AF_{PM} denotes the auditory filters of the Patterson-Moore model. Substituting (7.25) into (7.26) gives,

$$p=\frac{4f_0}{24.7(4.37f_0+1)}. \tag{7.28}$$

Substituting (7.28) into (7.27) gives,

$$AF_{PM}(f,f_0)=\left(1+\frac{4|f-f_0|}{24.7(4.37f_0+1)}\right)e^{\left[\frac{4|f-f_0|}{24.7(4.37f_0+1)}\right]}. \tag{7.29}$$

Equation (7.29) allows the response of the auditory filter centered at f_0 to be calculated as a function of frequency f (in kHz).

With a slight modification to (7.29), it is possible to calculate the excitation pattern across frequency due to a signal at frequency f_c.

$$\Im_{PM}(f_c,f)=\left(1+\frac{4|f_c-f|}{24.7(4.37f+1)}\right)e^{\left[\frac{4|f_c-f|}{24.7(4.37f+1)}\right]}. \tag{7.30}$$

Equation (7.30) can be used to determine the excitation level at frequency f due to a signal at f_c.

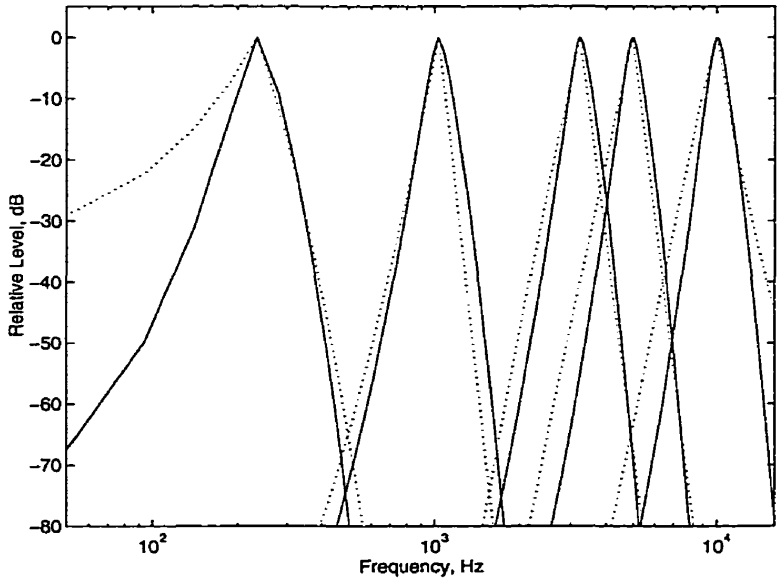


Figure 7.13 Comparison of predicted auditory filter responses. Solid curves: Patterson-Moore model, dotted curves: Zwicker model.

Equations (7.29) and (7.30) are now used to compare the Patterson-Moore model to the high resolution Zwicker model described earlier. Figure 7.13 compares the auditory

filters predicted by the Patterson-Moore model (solid curves) to those predicted by the Zwicker model (dotted curves) for an input signal with a level of 40 dB. It can be seen from the figure that significant differences exist between the two models. This is particularly true at low frequencies where the bandwidths of the Zwicker auditory filters are significantly wider. This result could be anticipated from Figure 7.12, where it was seen that the Zwicker estimate of the critical bandwidth was much larger than the predicted ERB's below 500 Hz. It should also be noted that the differences between the two models is not consistent across frequency. That is, at some frequencies the Zwicker auditory filters are wider than the Patterson-Moore auditory filters, whereas at other frequencies they are narrower. To better examine the differences in the auditory filters, they are presented on a linear frequency scale at selected frequencies in Figure 7.14. The differences in the auditory filters are more obvious in this figure, where differences of more than 10 dB can be found.

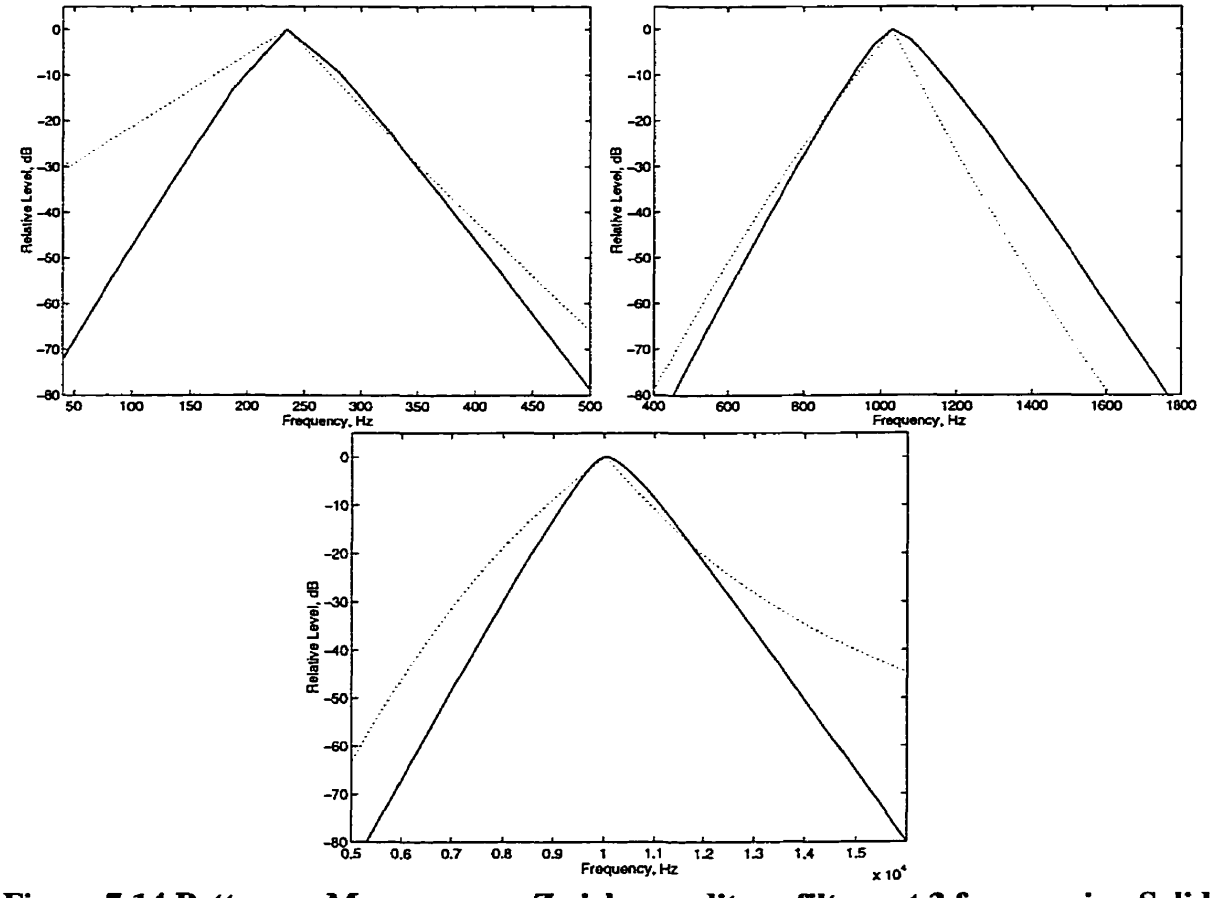


Figure 7.14 Patterson-Moore versus Zwicker auditory filters at 3 frequencies. Solid curves: Patterson-Moore model, dotted curves: Zwicker model.

Another way to compare the two models is in terms of the predicted excitation patterns. Figure 7.15 compares the excitation patterns predicted by the Patterson-Moore model (solid curves) to those predicted by the high resolution Zwicker model (dotted curves) for an input signal with a level of 40 dB. Again there are significant differences in the two models across frequency, with the largest differences occurring at the lower frequencies. The differences in the two predictions are also very large at the highest frequencies, but this may be of little practical concern.

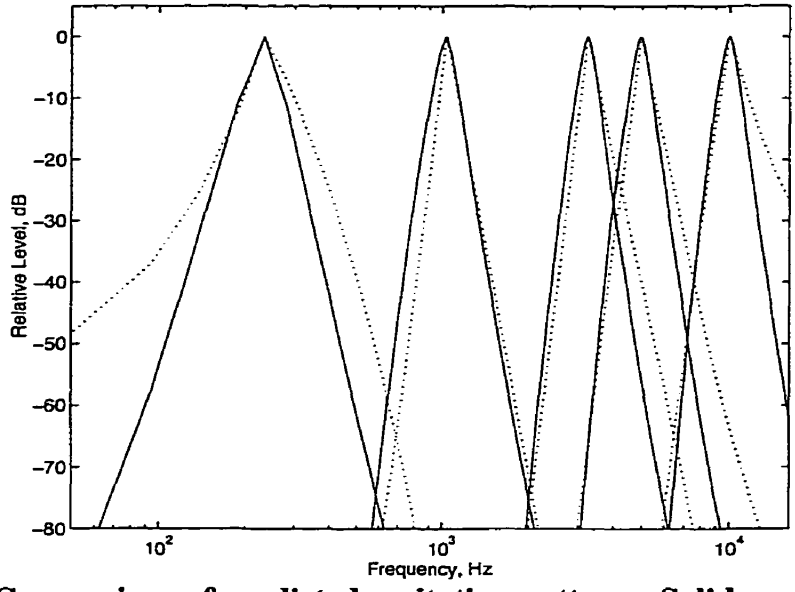


Figure 7.15 Comparison of predicted excitation patterns. Solid curves: Patterson-Moore model, dotted curves: Zwicker model.

Figure 7.16 provides a closer view of the excitation patterns on a linear frequency scale for three input signal frequencies. It can be seen that differences of more than 10 dB can occur in the excitation patterns predicted by the two models. Moreover, the differences are not consistent across frequency.

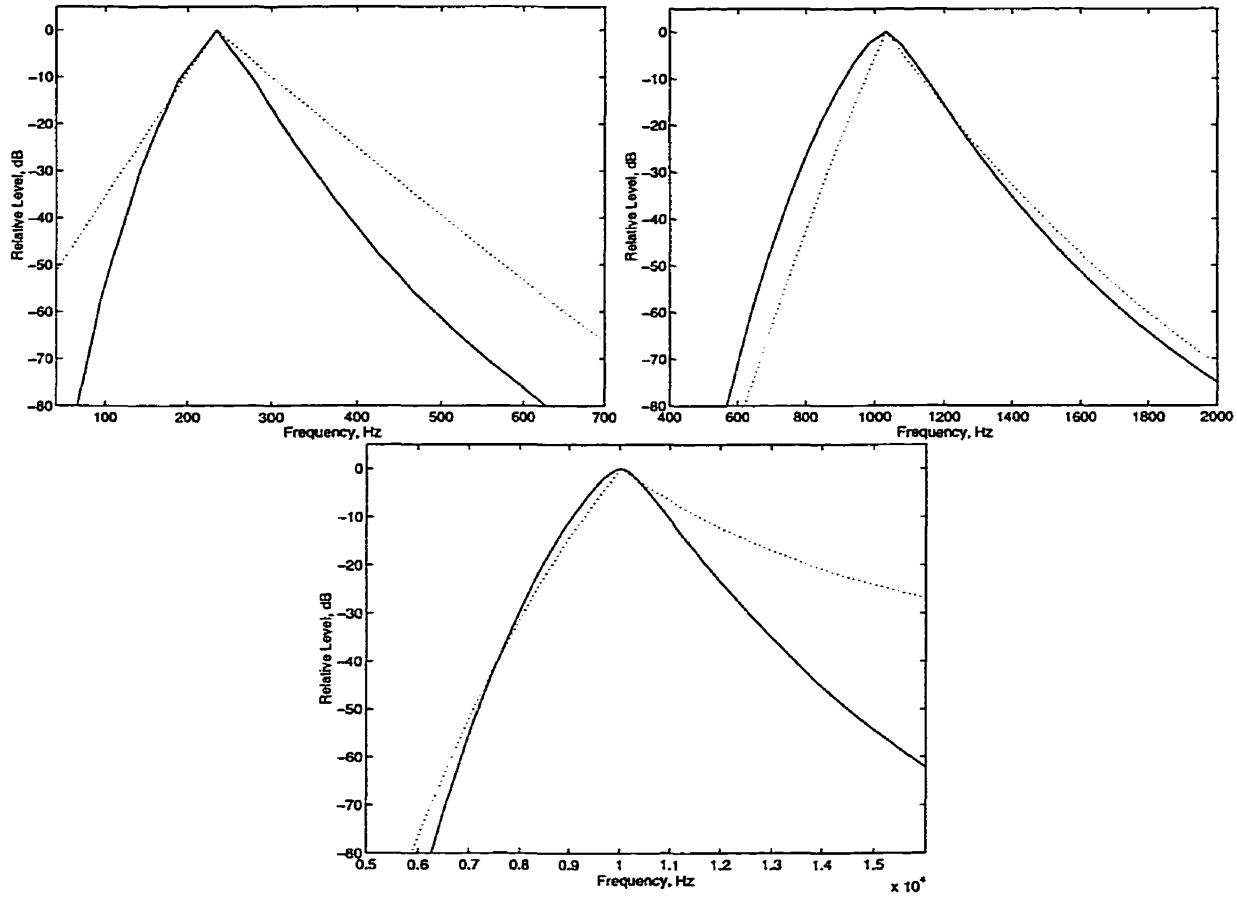


Figure 7.16 Patterson-Moore versus Zwicker excitation patterns at three frequencies. Solid curves: Patterson-Moore model, dotted curves: Zwicker model.

The figures comparing the auditory filters as well as the excitation patterns clearly indicate that, while they are both intended to predict the masking characteristics of the human auditory system, the two models produce significantly different predictions. Thus it is important to compare the performance of the two models for the noise reduction application.

7.3.3 Variations in the Shapes of the Auditory Filters with Level

Moore and Glasberg [156] measured the shapes of the auditory filters as a function of the level of the masker. They found that the slopes of the low-frequency skirts of the filters broadened as the level of the masker was increased. Conversely, the slopes of the high-frequency skirts of the auditory filters increased as the level of the masker was increased. In a later study, Glasberg and Moore [152] re-analyzed these results and concluded that, while the lower slopes do indeed vary with level, the upper slopes remain relatively unchanged with increasing masker level. This result is in keeping with the well known psy-

choacoustic phenomenon called upward spread of masking, wherein the masking threshold at frequencies above the masker increase nonlinearly with level [128,126].

Glasberg and Moore also found that the auditory filter centered at 1 kHz is approximately symmetrical when the level of the masker is 51 dB/ERB. The auditory filters at other centre frequencies were found to be symmetrical when the input levels to the filters (i.e., after the outer and middle ear attenuation) are equal to the level of 51 dB/ERB at 1 kHz. Glasberg and Moore propose the following expression to determine the parameter p_l of the low-frequency skirt of a filter as a function of input level,

$$p_{l(X)} = p_{l(51)} - 0.38(p_{l(51)}/p_{l(51,1k)})(X-51)\,, \tag{7.31}$$

where $p_{l(51)}$ is the value of p at the centre frequency for an equivalent noise level of 51 dB/ERB and $p_{l(51,1k)}$ is the value of p_l at 1 kHz for a noise level of 51 dB/ERB. The parameter X denotes the equivalent input noise level in dB/ERB. The value of $p_{l(51)}$ can be found using (7.22).

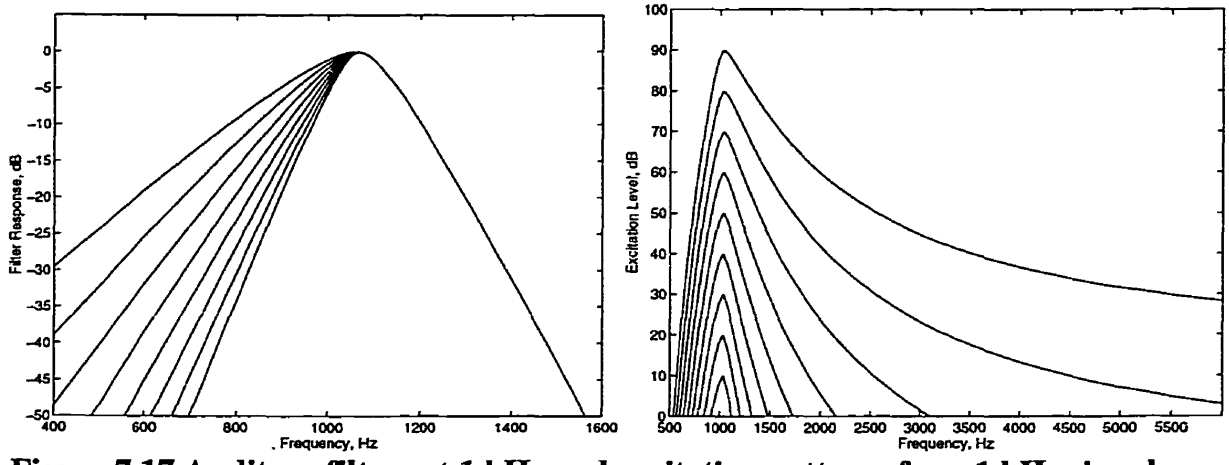


Figure 7.17 Auditory filters at 1 kHz and excitation patterns for a 1 kHz signal as a function of level.

The auditory filters at 1 kHz as a function of level (20 to 90 dBSPL) predicted by the Patterson-Moore model are shown in the left panel of Figure 7.17. It can be seen that the slope of the low-frequency skirt of the filter varies significantly. The steepest slope corresponds to an input level of 20 dB SPL while the shallowest slope corresponds to a level of 90 dBSPL. The right panel of the figure provides the excitation pattern as a function of level for 1 kHz sinusoidal input. It can be seen that the level of excitation at frequencies above 1 kHz is quite large, while for frequencies below 1 kHz, the level of excitation drops off very quickly. Stated differently, the signal produces a significant amount of masking above the input frequency, and much less masking below the input frequency.

The nonlinear relation between the level of the input signal and the amount of masking can also be seen from the figure. Higher level input signals result in a greater amount of masking above the frequency of the input.

It is of interest to compare the excitation patterns predicted by the Patterson-Moore model (solid curves) and the Zwicker model (dotted curves) as a function of level. This is done in Figure 7.18 using the Terhardt *et al.* expression (equation (5.109)) to account for the effects of level. The first panel in Figure 7.18 compares the predicted excitation patterns for a 250 Hz sinusoidal input signal. It can be seen that the excitation patterns predicted by the two models are very different in this frequency range. The Patterson-Moore model predicts much lower excitation levels than the Zwicker model. This is true for frequencies above and below the signal frequency. Differences in excess of 10 dB are evident in the upper frequencies, while differences in excess of 30 dB can be found in the lower frequencies. The discrepancies between the two models tend to increase for decreasing input levels.

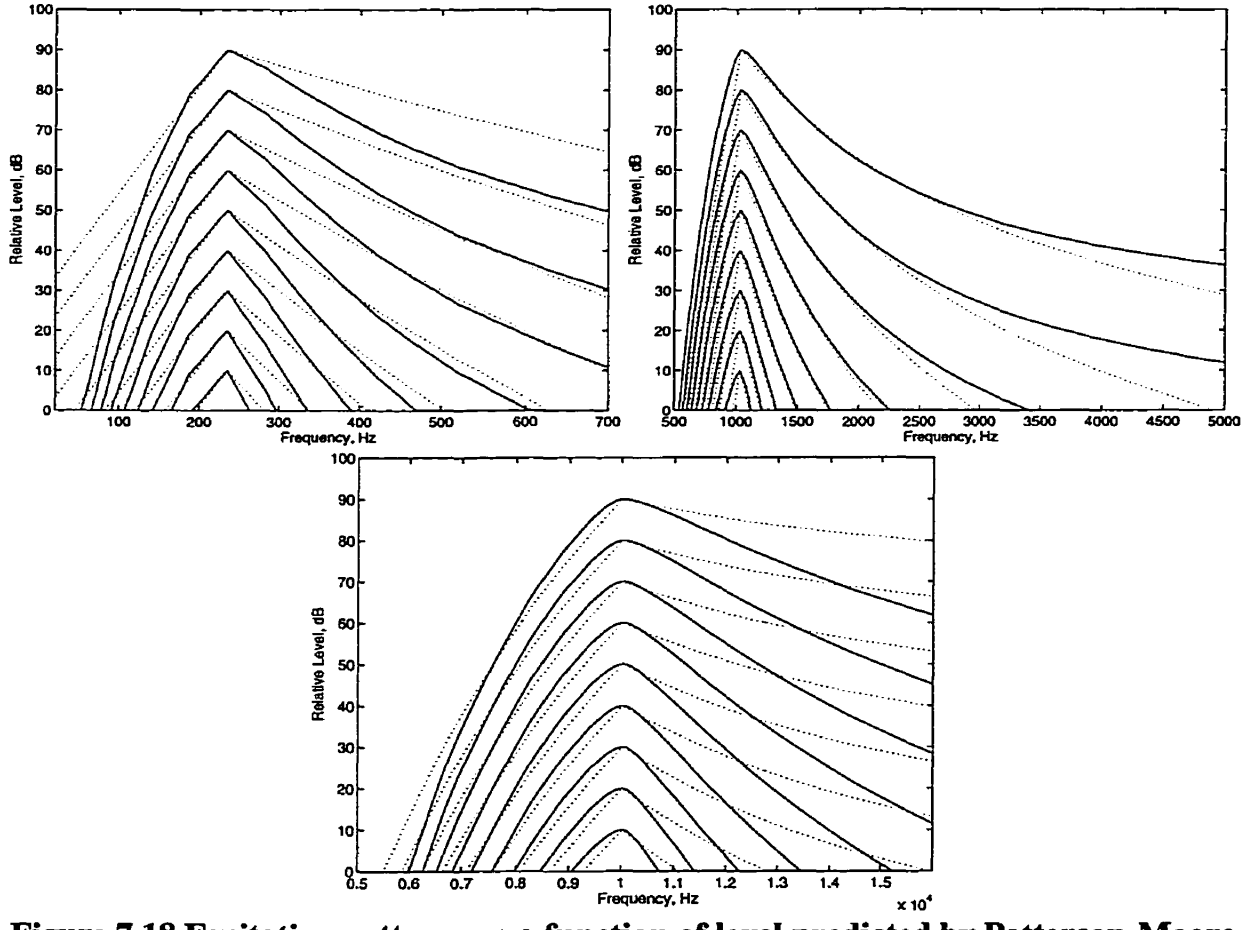


Figure 7.18 Excitation patterns as a function of level predicted by Patterson-Moore model (solid) and Zwicker model (dotted) using equation (7.11).

The second panel of the figure compares the predicted excitation for a 1 kHz sinusoidal input. The differences between the two models is much less in this frequency range. However, large differences do exist in the lower frequencies, where the Zwicker model predicts significantly lower excitation levels.

The third panel of the figure compares the predicted excitation for a 10 kHz sinusoidal input. In this frequency range, the two models are in reasonable agreement for the lower frequencies. However, at higher frequencies, the Zwicker model predicts significantly higher excitation levels than the Patterson-Moore model. The differences between the two models is in excess of 10 dB and increases for lower level signals.

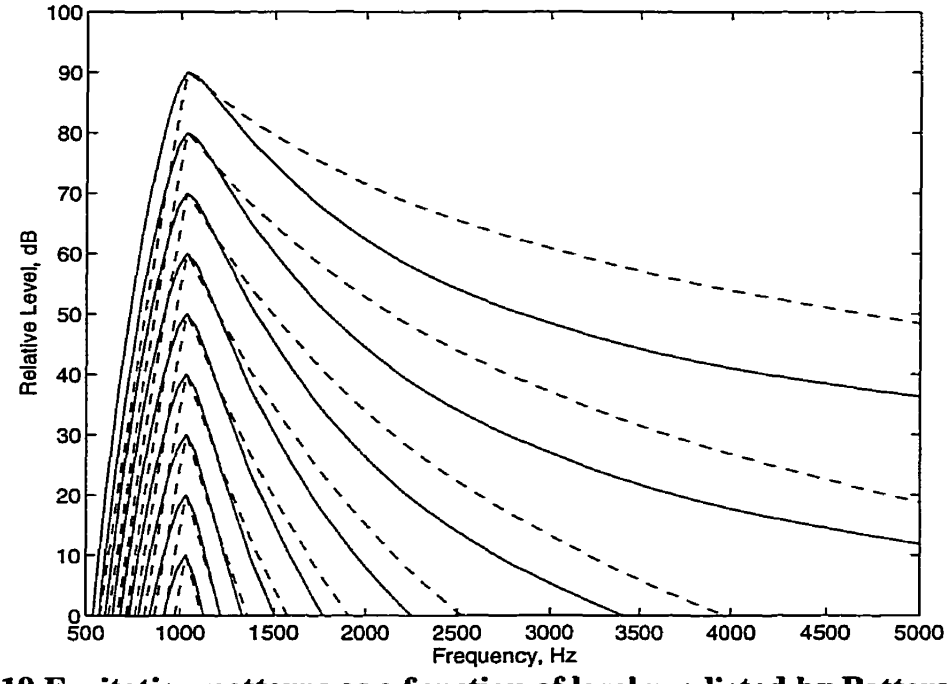


Figure 7.19 Excitation patterns as a function of level predicted by Patterson-Moore model (solid) and Zwicker model (dashed) using equation (7.10).

The excitation patterns shown in Figure 7.18 used the expression proposed by Terhardt *et al.* to account for the effects of level. As seen earlier, a different expression for level is used in PAQM and hence by Tsoukalas *et al.* (see equation (7.10)). Figure 7.19 compares the excitation patterns predicted by the Patterson-Moore model and the Zwicker model (using equation (7.10)) as a function of level. The excitation patterns are for a 1 kHz sinusoidal input. It is evident from the figure that the two models predict very different excitation patterns when the PAQM expression for level is employed. A comparison of Figure 7.19 and the second panel of Figure 7.18 shows that the use of this expression for level results in a much larger discrepancy between the Patterson-Moore model

and the Zwicker model. This discrepancy between the two models when using the PAQM expression for level is also larger for input signals at other frequencies.

The results of the comparison of the Patterson-Moore model and the high resolution Zwicker model clearly demonstrate that the two models predict significantly different excitation patterns. The differences are amplified when the expression for level used by Tsoukalas *et al.* is employed. The accuracy of the perceptual model is expected to be critical to the performance of a perceptually based spectral subtraction algorithm, and so the two models should be implemented and compared.

7.3.4 Non-Simultaneous Masking

The peripheral auditory models described so far are intended to predict the simultaneous component of masking. That is, the masker and the signal were assumed to occur at the same instant in time. In this section the non-simultaneous component of masking is explored. In non-simultaneous masking, the masker occurs either before (forward masking) or after (backward masking) the signal. Forward masking is the dominant of the two forms of non-simultaneous masking and can provide significant amounts of masking for up to 200 ms after the masker has terminated. Backward masking provides much less masking and is much less consistent across listeners. Also, it has been demonstrated that trained listeners can exhibit almost no backward masking [126,163]. As such, we will only consider forward masking in our perceptual model.

Forward masking has been studied by many researchers and certain consistent findings emerge. The amount of forward masking is greater for signals arriving nearer in time to the masker. Roughly speaking, the amount of forward masking is logarithmically related to the delay between the signal and the masker. The amount of forward masking decays to zero for delays longer than about 200 ms. Also, forward masking varies with frequency and level. However, the amount of forward masking is nonlinearly related to the level of the masker.

An analytic expression describing non-simultaneous masking as a function of level and frequency does not appear to be available in the literature. Therefore, we choose to derive such an expression based on the measured results from a few studies. The results of a study by Jesteadt *et al.* [164] are in good agreement with a later study by Moore and Glasberg [165] and will form the basis of our derivation.

Jesteadt *et al.* investigated the forward masking of a sinusoidal signal by another sinusoid of the same frequency [164]. They conducted measurements for frequencies ranging

from 125 to 4000 Hz. The level of the masker in their experiments was systematically varied in the range from 20 to 90 dBSPL. They found that an expression of the following form provided a good fit to their results at a given frequency,

$$FM(L_m) = a(b - \log \Delta t)(L_m - c) \text{ dB SL}, \tag{7.32}$$

where a, b, and c are parameters which must be fit to the data for a given frequency, Δt is the length of the delay between the masker and the signal, and L_m is the level of the masker measured in terms of the sensation level (SL). Jesteadt *et al.* provide values for the parameters a, b, and c for frequencies of 125, 250, 500, 1000 and 4000 Hz. The form of equation (7.32) is somewhat inconvenient for the present application since it does not provide straightforward estimates of the forward masking at arbitrary frequencies and levels. Also, the amount of predicted forward masking is not given in dBSPL. It is therefore desirable to derive an analytic expression which addresses these issues.

Using the values for the signal and masker thresholds provided by Jesteadt *et al.*, the forward masking thresholds were calculated from the data. To simplify the process, the analysis was restricted to delays of Δt =20 ms. This value of Δt was chosen since it corresponds to the typical time between processing frames in the spectral subtraction algorithm, assuming 50% overlap of the frames. A variety of analytic expressions were investigated to find one which provided a reasonable fit to the measured data at each masker level. The following expression was found to give a very good fit to the data,

$$FM(f) = \alpha + \beta e^{\left[\frac{-f}{\gamma}\right]} \text{ dB}. \tag{7.33}$$

The values in Table 7-1 for α, β, and γ were found to give the optimal fit at each masker level.

The values in Table 7-1 can be used in conjunction with equation (7.33) to calculate the forward masking for any frequency at a given masker level. For frequencies below 100 Hz, the above expression predicts excessive amounts of forward masking and the value predicted at 100 Hz should be used.

To form a single analytic expression for predicting forward masking, the parameters in Table 7-1 were fitted with appropriate equations. These equations were then substituted into equation (7.33) to yield the following expression,

$$FM(f,L)=(-221.71+24.57L)^{0.5}+(4.738+0.053L^2)e^{\left[\frac{-f(1-0.049L+0.0007L^2-1.659x10^{-6}L^3)}{117.14-6.03L+0.0846L^2}\right]}$$

dB, (7.34)

with 100 Hz $\leq f \leq$ 20 kHz and 10 dBSPL $\leq L \leq$ 100 dBSPL

where f is the frequency in Hertz and L is the level of the masker in dBSPL. For masker levels L below 10 dBSPL, a value of zero should be assigned to $FM(f,L)$. Again, it should be stated that equation (7.34) is only intended to predict the amount of forward masking for a delay of 20 ms. The expression developed above provides a very good prediction of the measured results of Jesteadt *et al.* as shown in Figure 7.20.

Masker Level dBSPL	α	β	γ
90	45.58	47.60	247.94
80	42.11	42.88	255.15
70	36.64	38.25	263.40
60	35.18	33.73	272.82
50	31.63	24.77	334.19
40	28.43	15.25	292.37
30	25.05	9.23	109.70
20	15.00	9.11	112.59
10	5.00	9.11	112.58

Table 7-1 Parameter values for equation 7.33 to predict forward masking.

The horizontal axis provides the measured data (for Δt = 20 ms) taken from the study by Jesteadt *et al.*, while the vertical axis shows the amount of forward masking predicted by (7.34). The forward masking predicted by (7.34) is in good agreement with the measured data and satisfies the requirement for a single analytic expression to predict forward masking.

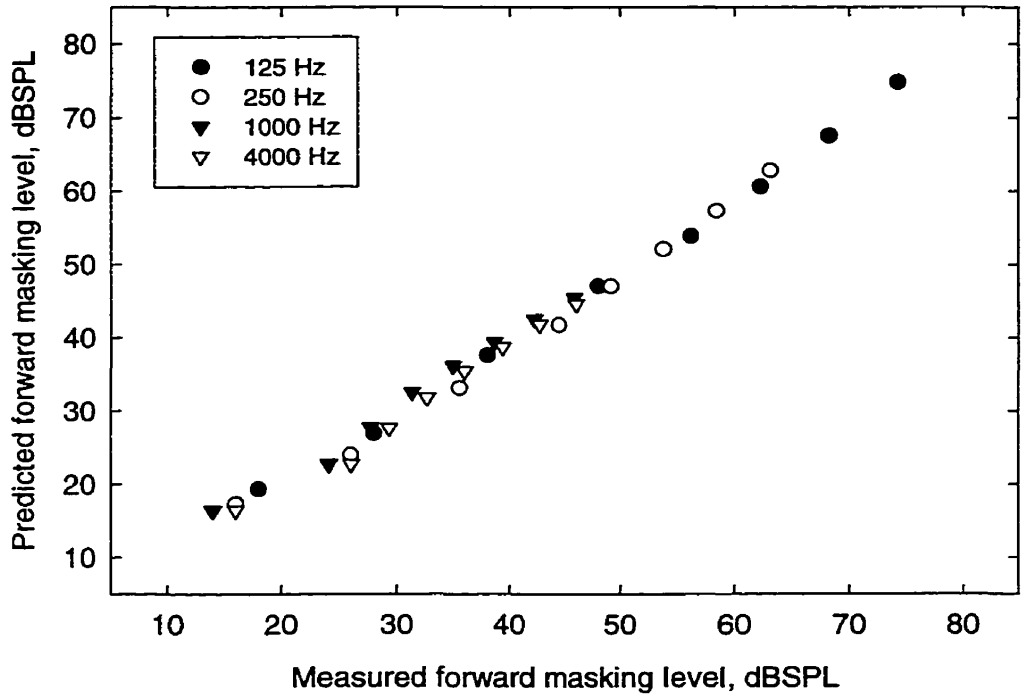


Figure 7.20 Forward masking predicted by equation (7.34) versus measured values.

7.3.5 Addition of Masking

In the previous sections, methods for predicting the amount of masking due to a single masker were described for both the simultaneous and non-simultaneous masking cases. The question arises of how to combine the masking effects due to two or more maskers. In the traditional power spectrum model of masking, the output of each auditory filter is simply the linear sum of the power of each masker component applied to the filter [124,166]. According to this model, if two maskers provide the same amount of masking individually, then the two maskers together should provide 3 dB more masking than either masker alone. However, results of experiments to measure the addition of masking indicate that this simple linear model is not appropriate.

Green [146] measured the masking thresholds of a signal for two independent simultaneous maskers, and compared these results to the masking threshold with the two maskers combined. The two maskers were chosen such that each provided the same amount of masking. Therefore, combining the two maskers should have increased the masking threshold by 3 dB. However, Green found that the combined masker produced 6 to 14 dB more masking than either masker alone. That is, Green found an *excess masking* of between 3 and 11 dB.

Lutfi [144] further investigated the issue of additivity of simultaneous masking by measuring the amount of excess masking produced by a variety of masker types and obtained between 10 and 17 dB of excess masking. Based on the work of Penner and Shiffrin [143], Lutfi proposed the following transformation to predict the combined masking of two simultaneous maskers

$$M_{ab} = \left[M_a^p + M_b^p\right]^{1/p}, \tag{7.35}$$

where M_a and M_b are the individual masking effects of the maskers and M_{ab} is the combined masking effect. The parameter p provides a compressive nonlinearity to account for the excess masking. Lutfi found that values for p between 0.20 and 0.33 provided the best fit to his measured data. This model of the addition of masking is referred to here as the power-law model.

In a later paper, Lutfi [145] reexamined his data as well as the data from studies by Canahl [167], Nelson [168], Zwicker [127], Green [146], Patterson and Nimmo-Smith [169], Bilger [170], and Moore [166] to see how well the power-law model could predict these results. He found that setting p to 0.33 provided a good fit to the data from each of these studies. In all cases, the power-law model with $p = 0.33$ was a far better predictor of the excess masking than the simple linear model.

Moore [166] voiced his skepticism of Lutfi's findings as well as the overall concept of excess masking, and devised a series of experiments wherein excess masking appeared to be either non-existent or far less than measured in previous studies. Moore suggested that the experimental procedure used by Lutfi (and others) did not account for off-frequency listening and thus led him to erroneous conclusions regarding excess masking in the simultaneous case. Moore included a broadband background noise in his measurements to limit the effect of off-frequency listening. Moore concluded that Lutfi's power-law model clearly fails in some situations and that the traditional linear model should not be abandoned.

In a more recent study, Humes and Jesteadt [147] reexamined Lutfi's results in the context of Moore's comments and rejected Moore's conclusions. Humes and Jesteadt contend that, in his experiments, Moore did not account for the effects of intermasker suppression, wherein one masker can suppress the effects of another. Moreover, Humes and Jesteadt demonstrated that the background noise used by Moore gave additional masking which was not accounted for. Humes and Jesteadt also argue that the internal noise of the peripheral auditory system must be considered as an additional masker which

is always present. They devised a modified power-law for predicting the addition of simultaneous masking which accounts for the internal noise. The modified power-law can be expressed as

$$M = 10\log\left\{\left[\left(\sum_{i=1}^{N}\left[10^{M_i/10}\right]^p\right)-\left[10^{ATH/10}\right]^p\right]^{1/p}\right\} \text{ dB}, \tag{7.36}$$

where M_i i=1,2,...,N are the levels of the various masking components, ATH is the absolute threshold of hearing at the signal frequency, and p is compression factor. Humes and Jesteadt found that setting the parameter p to 0.3 gave a very good fit to the results of Lutfi, as well as the results of Moore. Therefore the modified power-law appears to be the most appropriate model for predicting the addition of simultaneous maskers. It should be noted that for larger values of M_i, the modified power-law model is equivalent to the power-law model and so it is not surprising that the optimal value for p found by Humes and Jesteadt is very close to the value found by Lutfi.

Penner and Shiffrin [142,143] have studied the additivity of masking in the non-simultaneous case by comparing the masking due to individual and combined maskers. They found that excess masking also occurs in the non-simultaneous case, and that the amount of excess masking obtained is dependent on level. Higher masker levels provide greater amounts of excess masking. Penner and Shiffrin measured as much as 10 dB of excess masking in some of their test conditions. More recently, Oxenham and Moore [163] investigated the additivity of masking in the non-simultaneous case and obtained results which were in good agreement with those of Penner and Shiffrin.

To predict the masking due to two non-simultaneous maskers, Penner and Shiffrin proposed a high-compression model based on the following expression

$$J(M_{ab}) = J(M_a) + J(M_b) \tag{7.37}$$

$$J(M_x)=\left\{1+[(\beta-1)/10\log\alpha]M_x\right\}^{\log 2/\log\beta} \quad \alpha > 1.0 \tag{7.38}$$

where the parameters α and β are obtained by fitting the equation to the data. Humes and Jesteadt showed that their modified power-law (equation (7.36) above) provides a very good fit to Penner and Shiffrin's measurements. In the case of two sequential forward maskers, a value for p of 0.08 gave the best fit to the data, while a value for p of 0.23 was best when combining forward and backward maskers. Humes and Jesteadt also showed that the modified power-law performed well at predicting the results from other studies of non-simultaneous masking conducted by Wilson and Carhart [171], and Widin and Viemeister [172]. Therefore, it appears that the modified power-law, with the appropriate

choice of p, provides a good model for predicting the additivity of masking in both the simultaneous and non-simultaneous case.

In order to better appreciate the amount of excess masking that might be obtained as a result of two maskers, it is instructive to examine a plot of masking predicted by the original power-law model.

Figure 7.21 plots the amount of masking predicted by the power-law model for two maskers. In the figure one of the maskers is assumed to be at a constant level, and the abscissa indicates the level of the other masker relative to the first. Therefore, a value of 0 dB on the abscissa indicates that the two maskers are providing the same amount of masking. The ordinate shows the increase in masking due to the addition of the second masker. Curves are provided for three values of the compression factor.

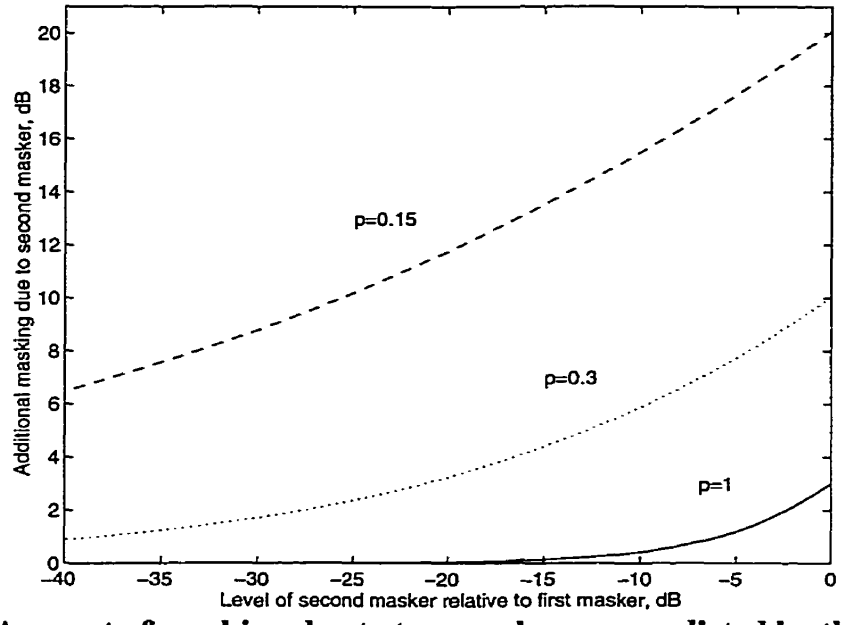


Figure 7.21 Amount of masking due to two maskers as predicted by the power-law with compression factor as a parameter.

With p=1 the power-law model is equivalent to the linear model. Therefore, when the two maskers provide equivalent amounts of masking, the total masking increases by 3 dB. With p=0.3, the power-law model predicts that the sum of the two equal maskers is 10 dB greater than the masking due to a single masker. That is, the model predicts 7 dB of excess masking. Also, with p=0.3, the second masker provides significant additional masking (about 3 dB) even when it's individual masking effect is 20 dB below the first. With p=0.15, the power-law model predicts that two equivalent maskers will provide 17 dB (20 - 3 dB) of excess masking. Furthermore, for this value of p, the second masker

provides a significant amount of additional masking even when it's individual masking effect is more than 40 dB below the other.

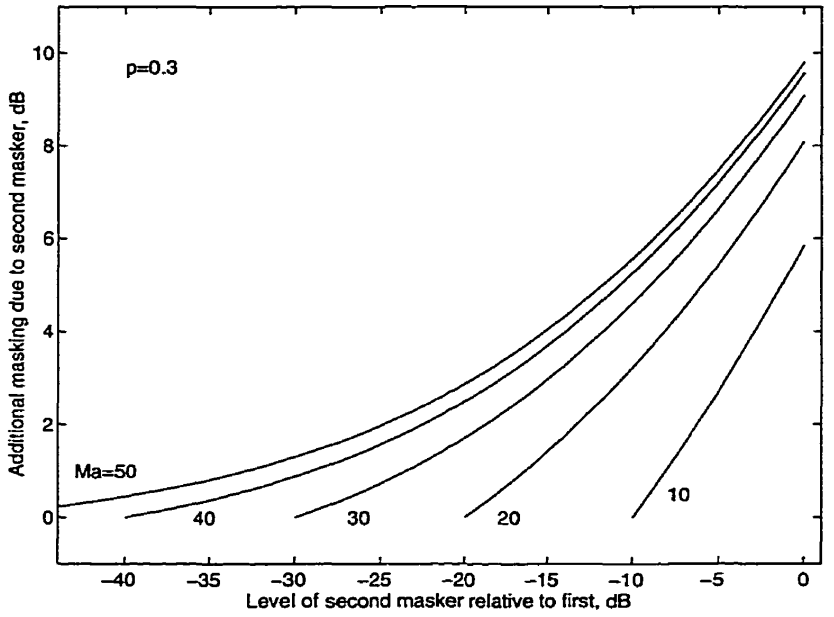


Figure 7.22 Amount of masking due to two simultaneous maskers as predicted by the modified power law with masker level as a parameter.

Figure 7.22 shows the amount of combined masking due to two simultaneous maskers predicted by the modified power-law model with p=0.3. The various curves in the figure indicate the amount of masking due to the more dominant masker. That is, the amount of masking predicted by the modified power-law model is dependent on the level of the masker. Therefore, if the two maskers each provide 10 dB of masking, then their combined masking is 16 dB (10 + 6 dB) according to the model. At the other extreme, if the two maskers each provide 50 dB of masking, then their combined masking is 60 dB (50 + 10 dB) according to the model.

Figure 7.23 shows the amount of combined masking due to two non-simultaneous maskers predicted by the modified power-law model with p=0.08. Again, the various curves in the figure indicate the amount of masking due to the more dominant masker. Due to the lower value of p, the excess masking found in the non-simultaneous case is greater than was seen in the simultaneous case. This is particularly true at higher masker levels. Here, if the two maskers each provide 10 dB of masking, then their combined masking is 18 dB (10 + 8 dB) according to the model. At the other extreme, if the two maskers each provide 50 dB of masking, then their combined masking is 76 dB (50 + 26 dB) according to the model. The effect of the compression factor p and the importance of accounting for excess masking is thus evident from these figures.

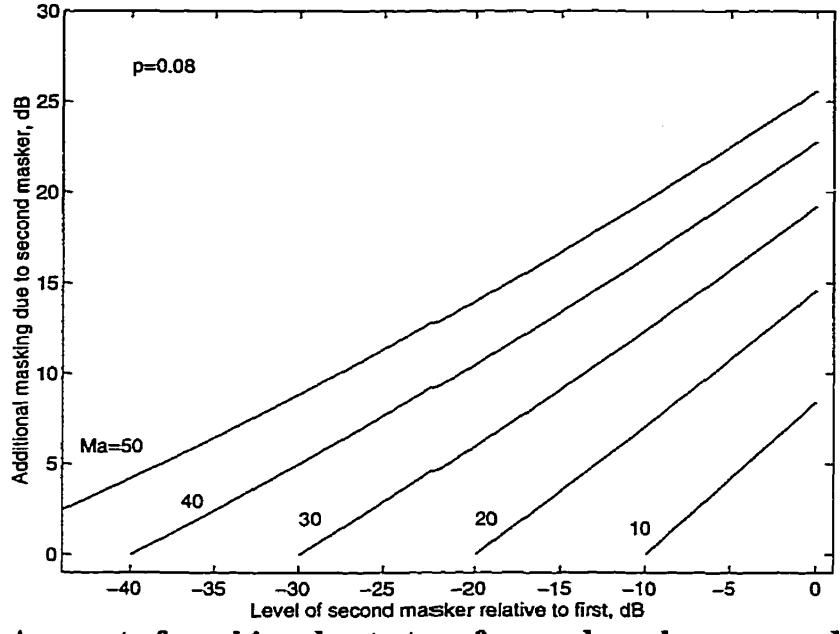


Figure 7.23 Amount of masking due to two forward maskers as predicted by the modified power law with masker level as a parameter.

7.4 Applying Perceptual Models to Spectral Subtraction

In the previous sections simultaneous masking predicted by the high resolution Zwicker model as well as the Patterson-Moore perceptual model were examined. It was seen that while the two models are both intended to predict the auditory masking threshold for a given input signal, their predictions are quite different. In particular, the responses of the level dependent auditory filters were shown to be significantly different for the two models. In this section a generalized approach for implementing the various components of the perceptual model is given. The perceptual model is then integrated with the spectral subtraction algorithm. This allows the performance of the two perceptual models to be compared.

Figure 7.24 shows a general block diagram of a perceptual model which includes each of the components described in this chapter. The model accepts the measured spectral magnitude of a signal at its input, and provides an estimate of the masking threshold for that signal. It should be noted that the components of the auditory model relating to non-simultaneous masking and the addition of masking are not inherent to either of the two models. Rather, they were derived as part of the present study. Also, the expressions used in the Patterson-Moore model to describe the outer and middle ear transfer functions, as well as the internal noise floor were derived in the present study.

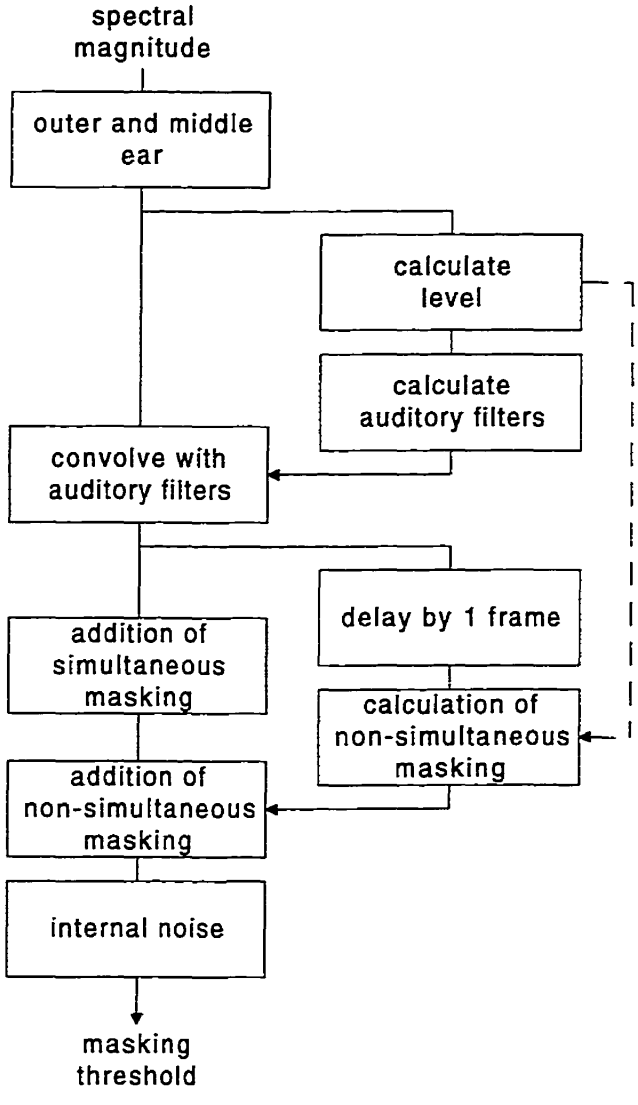


Figure 7.24 Block diagram of perceptual model.

In order to compare the two perceptual models, they were each incorporated into the subband/sub-frame spectral subtraction algorithm. Figure 7.25 shows the spectral subtraction process with the inclusion of a perceptual model. As shown in the figure, the model is used to provide a perceptual based estimate of the noise. As in the traditional spectral subtraction process, this noise estimate is subtracted from the input signal.

In Figure 6.14, it was seen that in the subband/sub-frame based spectral subtraction algorithm, separate zero-phase spectral subtraction filters $\Psi_{ij}(k)$ are applied to each of the time-frequency cells $v_{ij}(k)$. These filters can now be replaced with perceptual based spectral subtraction filters. However, the perceptual based spectral subtraction filters are not entirely independent of each other. Some information regarding masking levels must be exchanged between the filters operating on the different subbands.

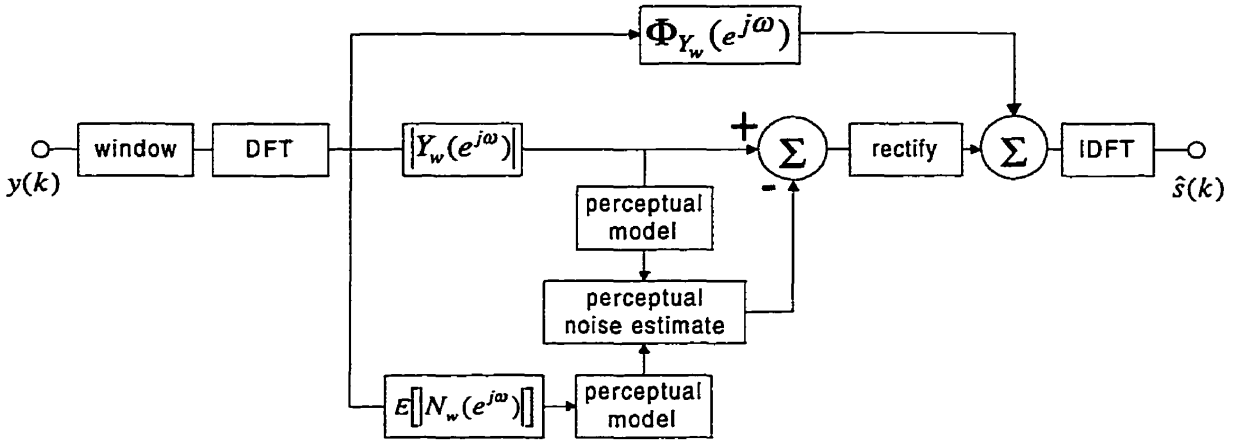


Figure 7.25 Perceptual based spectral subtraction.

In describing the simultaneous masking component of the auditory models, it was seen that a masker can mask signals that are both higher and lower in frequency. Therefore, it is necessary to exchange this masking information between subbands. Specifically, the energy in the lower frequency subbands will contribute to the masking threshold in the higher frequency subbands. As such, some estimate of this masking must be passed from the lower frequency subbands to the higher frequency subbands. Conversely, the energy in the upper frequency subbands will also provide a degree of masking (although much less) to the lower frequency subbands, and that information needs to be passed to the lower subbands.

7.4.1 Estimating the Clean Signal

In deriving the perceptual based spectral subtraction filter, it was assumed that the clean signal was available in order to determine the masking threshold. Of course, in a practical situation this is not the case. Therefore, the various means of obtaining an estimate of the clean signal must be considered. This matter has received very little attention in the literature.

Tsoukalas *et al.* [130] do not determine an estimate of the clean signal explicitly, but use the following equation to determine the zero-phase spectral subtraction filter $\Psi(\omega_n)$.

$$\Psi(\omega_n)=1-\frac{PM\{E[|N(\omega_n)|]\}}{PM\{|Y(\omega_n|)\}}, \tag{7.39}$$

where we use $PM\{\ \}$ to denote the processing of a signal through the perceptual model. It can be seen that, in this approach, the noise estimate which is made during periods without speech activity is processed through the perceptual model. This, in effect, provides an

estimate of the excitation pattern due to the noise. The denominator of the equation is the excitation pattern due to the noisy input signal which is determined on a frame-by-frame basis.

A key advantage of this approach is that the estimate of the excitation pattern due to the noise (i.e., the numerator) only needs to be calculated once, and does not need to be updated for each frame. This provides a significant reduction in computational complexity since a major component of the perceptual model involves multiplication of a 1 by N vector with an N by N matrix to account for the auditory filter bank.

A shortcoming of the above method is that it ignores the fact that the width of the auditory filters vary significantly with level. In the method described above, the noise is processed through the perceptual model in the absence of the signal and is therefore processed assuming a fixed sound pressure level. However, in reality, the noise is mixed with the desired signal and so the sound pressure level changes on a frame-by-frame basis. Therefore, the width of the auditory filters used in the perceptual model should also change for each frame. It should be noted that it is not the level of the noise that is changing on a frame-by-frame basis, but rather the overall level of the noisy signal. By not accounting for the change in the auditory filters with level, the method of Tsoukalas *et al.* also inherently ignores the nonlinear addition of masking that was seen earlier.

To overcome these shortcomings, the following method for calculating the zero-phase spectral subtraction filter is proposed

$$\Psi(\omega_n)=\frac{PM\{|\breve{S}(\omega_n)|\}}{PM\{|Y(\omega_n)|\}}=\frac{PM\{|Y(\omega_n)|-E[|N(\omega_n)|]\}}{PM\{|Y(\omega_n)|\}}. \tag{7.40}$$

In this approach, an initial estimate $|\breve{S}(\omega_n)|$ of the spectral magnitude of the clean signal is made using a traditional spectral subtraction algorithm. This estimate of the clean signal is then processed through the perceptual model PM yielding an estimate of the excitation pattern due to the clean signal. The excitation pattern due to the noisy signal is also found. The main advantage of this approach is that both excitation patterns are calculated using the correct auditory filters, and these calculations are updated with each frame. Also, this method allows complete flexibility in the choice of spectral subtraction parameters for determining the initial estimate of the clean signal. The disadvantage of this method is that two excitation patterns must be determined for each frame and thus the computational complexity of the algorithm is significantly increased.

Informal listening tests were conducted to evaluate the two approaches. It was found that the new method significantly outperforms the method proposed by Tsoukalas *et al.*

This method is therefore recommended for situations where the highest quality output is required.

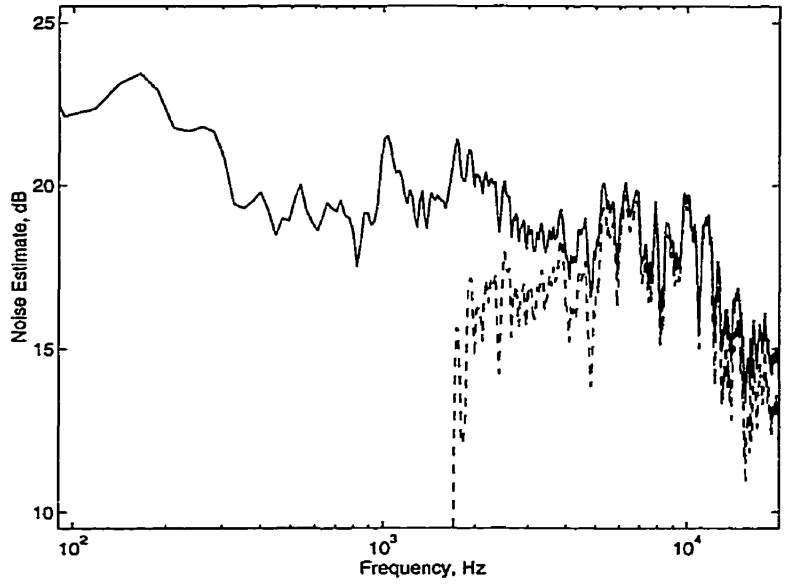


Figure 7.26 Noise estimates with and without a perceptual model.

With the perceptual model incorporated into the spectral subtraction algorithm, it is of interest to investigate how the model affects the processing applied to a noisy input signal. Figure 7.26 shows an example of the effect of the perceptual model on the noise estimate for a given processing frame. The solid curve in the figure shows the noise estimate without incorporating the perceptual model. The dashed curve shows the noise estimate when the perceptual model is used and the masking due to the speech signal is taken into account. As can be seen, for frequencies below about 2 kHz, the perceptual model predicts that the noise in this frame is entirely masked by the desired (speech) signal. Therefore, no spectral components are subtracted from the noisy signal at these frequencies. Above 2 kHz, the amount of masking due to the speech signal is reduced, and so a larger noise estimate is subtracted from the input signal. Therefore, by incorporating a perceptual model into the spectral subtraction process, the amount of overall processing applied to the input signal is dramatically reduced. A comparison of the performance of the spectral subtraction algorithm with and without a perceptual model is provided on the compact disc accompanying the thesis.

7.4.2 The Effect of Windows in Perceptual Based Spectral Subtraction

When describing the basic spectral subtraction algorithm in Chapter 5, it was stated that several researchers had investigated the effect of the type of window used prior to performing the DFT. They concluded that the choice of window did not have a significant

effect on the performance of the spectral subtraction algorithm. In this section, we examine how the choice of window can be important for perceptual based spectral subtraction algorithms.

In the traditional spectral subtraction method, a windowing function (e.g. Hanning) is applied to the input samples prior to performing the DFT. To account for the effect of the window, the input signal is processed in overlapping frames. After the signal has been enhanced in the frequency domain, it is transformed back to the time domain and the frames are overlapped when constructing the output signal. A rectangular window is applied when constructing the output signal.

Due to the temporal aliasing described in Section 6.2, the samples at the beginning and end of a processing frame may not decay to zero as they should. This can cause discontinuities at the boundaries of each frame which can result in audible clicks in the output signal. This is particularly true when the signal-to-noise ratio of the input signal is low and aggressive processing must be applied since this results in more severe temporal aliasing. The problem of discontinuities can also be more severe for subband/sub-frame based spectral subtraction since the amount of processing applied to the input signal can vary significantly from frame to frame. The problem of discontinuities can be seen as the vertical lines in Figure 5.7.

To resolve this problem we propose a method which is employed in perceptual based audio codecs where discontinuities at the boundaries of processing frames are a significant concern. To eliminate the discontinuities a synthesis windowing function is applied to the output signal. That is, an analysis windowing function is applied to the input signal, and a synthesis windowing function is applied to the output signal. The constraint described by (5.16) must now be generalized to allow for the synthesis window,

$$\sum_i w_{a_i}(k) \cdot w_{s_i}(k) = 1, \tag{7.41}$$

where $w_{a_i}(k)$and $w_{s_i}(k)$ are the analysis and synthesis windows respectively. Equation (7.41) states the well-known condition that the analysis/synthesis windows must sum to unity. It should be noted that the Hanning window or the Bartlett window recommended for traditional spectral subtraction do not satisfy the constraint of (7.41). Therefore, other windows must be investigated.

Typically, in perceptual audio codecs, a sine window is used for both the analysis and synthesis window [173,174]. The sine window satisfies (7.41) and provides a relatively

narrow main lobe and reasonable attenuation of the side lobes [175]. However the sine window has certain limitations when used in conjunction with a perceptual model. In the perceptual models, the convolution of the auditory filters with the magnitude spectrum of the signal provides an estimate of the masking threshold for the signal. This convolution effectively causes a smearing of the signal in the frequency domain. The windowing function applied to the input signal prior to performing the DFT also causes a form of smearing of the signal in the frequency domain. Therefore, the window causes excess smearing, beyond what is desired for the auditory model. This can be seen in Figure 7.27 and Figure 7.28. Figure 7.27 shows the excitation pattern for a 250 Hz input signal. The solid curve shows the response to the 250 Hz input signal as predicted by the Patterson-Moore auditory filters. The dotted curve in the figure shows the combined effect of a sine window and the Patterson-Moore auditory filters. Also included in the figure is the absolute threshold of hearing. The dashed curve shows the combined effect of the KBD window which will be described later.

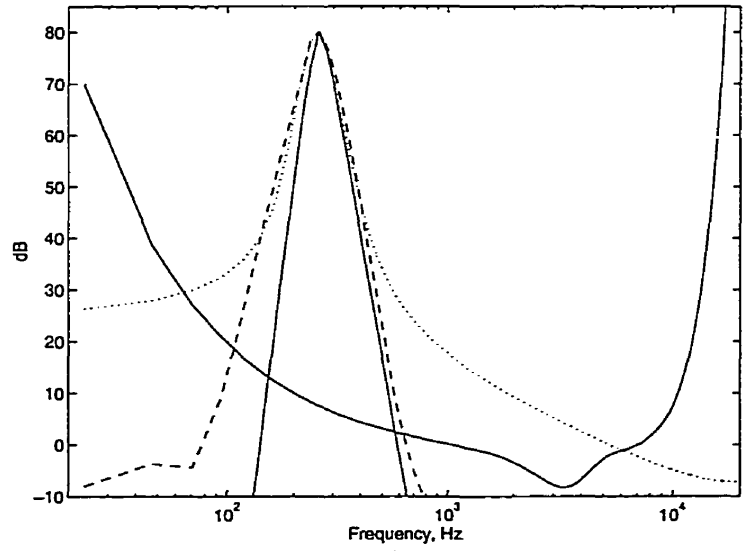


Figure 7.27 Effect of windows for a 250 Hz input signal. Solid curve: Patterson-Moore model, dotted curve: effect of sine window, dashed line: effect of KBD window.

It can be seen that the sine window causes the excitation pattern to be broader (in frequency) than desired. That is, the sine window causes the amount of masking due to the signal to be significantly overestimated. This same effect, although somewhat reduced, can also be seen in Figure 7.28 for a 1 kHz input signal.

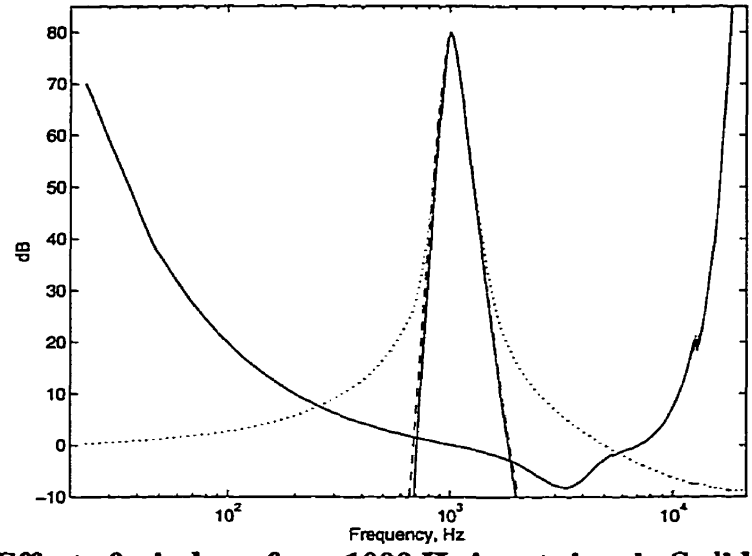


Figure 7.28 Effect of windows for a 1000 Hz input signal. Solid curve: Patterson-Moore model, dotted curve: effect of sine window, dashed line: effect of KBD window.

An alternative choice of windows is the Kaiser-Bessel Derived (KBD) window which was developed for the Dolby AC-3 audio codec [176,177]. The KBD window is also used in the newly developed MPEG AAC codec [174,178]. This window function was designed with the intent of providing a main lobe which is as narrow as possible, while attenuating the side lobes to a level below the threshold of hearing. The goal of the window is to minimize the number of bits needed to encode an audio signal. The KBD window is used as both the analysis and synthesis window.

The first step in creating the KBD window is to convolve a kernel window (the Kaiser-Bessel) with a rectangular window and the KDB window is obtained by taking the square root of the result. The KBD window is defined mathematically as follows,

$$w_a(k)=w_s(k)=\sqrt{\frac{\sum_{j=L}^{M} w(j)r(k-j)}{\sum_{j=0}^{K} w(j)}} \quad k=0,1,\ldots,N-1 \tag{7.42}$$

where

$$L=\begin{cases} 0 & 0\le k<N-K \\ k-N+K+1 & N-K\le k<N \end{cases}$$

$$M=\begin{cases} k & 0\le k<K \\ K & K\le k<N \end{cases}$$

In (7.42), $w(k)$ is the kernel window of length K+1, $r(k)$ is a rectangular window of length N-K, N is the size of the transform, and K is the width of the transition region of the KBD window. Within the Kaiser-Bessel kernel window there is a parameter α which allows a tradeoff between the width of the main lobe and the attenuation of the side lobes. The AC-3 codec uses $\alpha = 5$ [176], while the MPEG AAC codec can select between $\alpha = 4$ and $\alpha = 6$ [178]

The effect of the KBD window in conjunction with the smearing due to the auditory filters can be seen as the dashed curves in Figure 7.27 and Figure 7.28. It can be seen that the KBD window provides a significant improvement over the sine window. For an input signal of 250 Hz, the KBD window still causes the amount of masking at the lower frequencies to be overestimated somewhat. However, masking in the upper frequencies is very well predicted. At 1 kHz the KBD window causes almost no excess smearing and hence no overestimation of the masking threshold. Therefore, the KBD window is better suited to the task of predicting the masking threshold for a signal.

A further refinement to obtaining an accurate masking threshold can be had by modifying the slopes of the auditory filters to account for the additional spreading due to the KBD window. That is, the slopes of the auditory filters should be made steeper so that, once the KBD windowing function is included, the combination of the window and the auditory filters will provide the correct masking threshold. With the perceptual model based on auditory filters (i.e., the Patterson-Moore model), this is a relatively straightforward refinement. The equation describing the response of the auditory filters (see equation (7.14)) would then be

$$W(g)=(1+\tilde{p}g)e^{-\tilde{p}g} \qquad \tilde{p}\geq p, \tag{7.43}$$

where $\tilde{p}$ is a modified version of p which accounts for the effect of the chosen window.

As a final point regarding the choice of window in a perceptual based spectral subtraction algorithm, it is important to realize that the signal reaching the perceptual model does not include the effects of the synthesis window. That is, the perceptual model predicts the masking threshold using data which has only been windowed by the analysis window. Since the sum of the overlapping analysis windows does not equal unity (see (5.16)), there is an inherent error in the information received by the perceptual model. Depending on where it arrives within a processing frame, a signal component will be either emphasized or de-emphasized with respect to other components within the window.

To resolve this we propose the following. The input samples are windowed by an analysis window which satisfies the constraint of equation (5.16) prior to the DFT. This frequency domain information is then used by the perceptual model to derive the masking threshold for that frame. In parallel to this, the input samples are also windowed by the KBD analysis window (or a similar window) prior to performing a separate DFT. The noise reduction (spectral subtraction) is performed on this transformed data, but the perceptual noise estimate for that frame is calculated from the other transform data.

In order for this process to be successful, it is necessary to derive a window function which satisfies the constraint of (5.16) and has frequency characteristics which are similar to the response of the KDB window. The following window was found to satisfy both requirements,

$$w_a(k)=\frac{\sum_{j=L}^{M} w(j)r(k-j)}{\sum_{j=0}^{K} w(j)} \quad k=0,1,\ldots,N-1 \tag{7.44}$$

where

$$L=\begin{cases} 0 & 0\le k<N-K \\ k-N+K+1 & N-K\le k<N \end{cases}$$

$$M=\begin{cases} k & 0\le k<K \\ K & K\le k<N \end{cases}$$

In (7.44), $w(k)$ is a Kaiser-Bessel kernel window of length $K+1$, $r(k)$ is a rectangular window of length $N-K$, N is the size of the transform, and K is the width of the transition region of the new window. It is clear that this is very similar to the KBD window, except that the square-root has been removed in order to satisfy (5.16). In order to obtain a frequency response which is similar to the KBD window's response with $\alpha = 5$, a value of $\alpha = 2.5$ was used in the Kaiser-Bessel kernel of the new window.

The above discussion regarding the various aspects of choosing an appropriate window can be extended directly to the case of non-uniform sub-framing as depicted in Figure 6.19 and Figure 6.20. The refinements to the windowing process described in this section appear to have been overlooked by the developers of high quality perceptual audio coders.

7.5 Summary

In this chapter, the use of a perceptual model to improve the performance of the spectral subtraction algorithm was introduced. The psychoacoustic model based on critical bands was described and its limitations were highlighted. A new psychoacoustic model, based largely on the work of Patterson and Moore, was derived. A comparison between this model and a model based on the work of Zwicker revealed significant differences even though they are intended to predict the same masking threshold.

Based on the experimental results of Jesteadt *et al.*, an expression for predicting non-simultaneous masking as a function of level and frequency was derived. Expressions describing the nonlinear addition of simultaneous and non-simultaneous masking were also described.

A spectral subtraction process incorporating a perceptual model into the subband/sub-frame structure was described, and it was shown how the perceptual model can significantly reduce the amount of processing applied to the signal. Finally, the effects of the choice of window when using a perceptual based model was described and a method for reducing these effects was introduced.

The main objective behind the use of a perceptual model, as well as the use of subbands and sub-frames is to minimize the overall amount of processing applied to the noisy signal. By minimizing the amount of processing the severity of the residual artifacts and signal distortion is reduced. The subbands and sub-frames allow the processing to be directed by the characteristics of the noise, while the perceptual model allows the processing to be directed by the masking provided by the signal.

The work in this chapter which examined perceptual models, including a comparison of the simultaneous masking models, the development of new expressions for the outer and middle ear response, the derivation of an expression for non-simultaneous masking, and the examination of the effects of the window function, is of value to many other applications (e.g. perceptual audio codecs) not related to noise reduction.

8. EVALUATION OF THE NOISE REDUCTION ALGORITHMS

In this chapter the performance of the various noise reduction schemes described in this thesis are evaluated. Ideally, formal subjective listening tests would be conducted to evaluate and compare each of the various schemes. However, due to the complexity, as well as the time required to properly conduct such tests, it is not feasible to evaluate each scheme in this manner. Therefore, a formal subjective test was conducted to evaluate only the most promising noise reduction algorithms. The results of informal listening tests are used to describe the performance of the remaining noise reduction schemes. Also, a compact disc (described in this chapter) is included with the thesis to demonstrate the various noise reduction schemes as well as some of the artifacts which can result.

Before listening to the entire CD, the reader may wish to hear a few selected tracks in order to briefly compare the performance of the noise reduction algorithm developed in this thesis to the *classic* spectral subtraction algorithm. For this purpose, we suggest that the reader listen to the following three tracks.

Track #3: *Ref+12dB* Speech with camera noise at highest SNR used in the tests.
Track #24: *Ref+12dB* input signal (Track #3) processed using the Boll's method.
Track #33: *Ref+12dB* input signal (Track #3) processed using with subbands/sub-frames and the Patterson-Moore perceptual model.

Demo 1: Quick overview of algorithm performance.

The first 3 tracks on the CD contain a segment of speech which has been corrupted by camera noise. The level of the camera noise is varied in 6dB increments between the 3 tracks. Track #1 has the highest level of camera noise and is referred to as *Ref*. The level of the camera noise in Track #2 is 6dB lower (i.e., SNR is 6dB higher) and is referred to as *Ref+6dB*. The level of the camera noise in Track #2 is 12dB lower than Track #1 (i.e., SNR is 12dB higher) and is referred to as *Ref+12dB*. These 3 tracks are used throughout this chapter as the input signals to demonstrate the performance of the various noise reduction algorithms.

Track #1: *Ref* Speech with camera noise at lowest SNR.
Track #2: *Ref+6dB* Speech with camera noise at intermmediate SNR.
Track #3: *Ref+12dB* Speech with camera noise at highest SNR.

Demo 2: Reference input signals with camera noise.

8.1 Performance of Noise Reduction Techniques based on Adaptive Filtering Methods

Chapter 4 described noise reduction techniques based on adaptive filtering methods. To satisfy the requirement that a successful camera noise reduction scheme must be a single input system, an adaptive noise cancellation algorithm using a synthesized reference input signal was proposed. This algorithm was shown to provide a degree of reduction to the periodic component of the camera noise. It was also seen that, due to jitter in the arrival times of the camera noise pulses, the effectiveness of this noise reduction scheme was compromised. Therefore, steps were taken to synchronize the ANC process to the camera noise. This resulted in improved noise reduction.

The next six tracks on the CD demonstrate the performance of the ANC algorithm with a synthesized reference input. Track #4 is the noisy speech segment (Track #1) processed by the non-synchronized ANC system. Track #5 is the same noisy speech segment (Track #1) processed by a synchronized ANC system.

Track #4: *Ref* input signal (Track #1) processed by the non-synchronized ANC system.
Track #5: *Ref* input signal (Track #1) processed by the synchronized ANC system.

Demo 3: ANC with synthesized reference on lowest SNR input.

Track #6 is the noisy speech segment (Track #2) processed by the non-synchronized ANC system. Track #7 is the same noisy speech segment (Track #2) processed by a synchronized ANC system.

Track #6: *Ref+6dB* input signal (Track #2) processed by the non-synchronized ANC system.
Track #7: *Ref+6dB* input signal (Track #2) processed by the synchronized ANC system.

Demo 4: ANC with synthesized reference on intermediate SNR input.

Track #8 is the noisy speech segment (Track #3) processed by the non-synchronized ANC system. Track #9 is the same noisy speech segment (Track #3) processed by a synchronized ANC system.

Track #8: *Ref+12dB* input signal (Track #3) processed by the non-synchronized ANC system.
Track #9: *Ref+12dB* input signal (Track #3) processed by the synchronized ANC system.

Demo 5: ANC with synthesized reference on highest SNR input.

From the six tracks it is apparent that, as described in Chapter 4, the synchronized ANC system provides more noise reduction than the non-synchronized ANC system. Specifically, the synchronized ANC system provides more reduction at the higher frequencies. However, neither system provides complete elimination of the periodic component of the camera noise, and of course, neither system provides any reduction of the cyclical random component of the noise. Interestingly, by removing a portion of the periodic component of the noise, one can begin to hear the other components of the camera noise more clearly. Also, the perceived noise reduction is more dramatic when the input signal has a lower initial signal-to-noise ratio.

It can be noted on these tracks that the ANC methods do not provide much noise reduction during the intervals where there is speech activity. This makes it difficult to combine this method with the spectral subtraction methods since they assume that the noise is locally stationary. Efforts to combine the two approaches results in good noise reduction during the intervals without speech activity, but poor performance during the intervals where there is speech activity.

From these tracks it can be concluded that, as stated in Chapter 4, the ANC-based methods do not provide a sufficient degree of noise reduction on their own. Therefore, the ANC-based noise reduction methods were not included in the formal subjective test described later in this chapter.

8.2 Performance of Noise Reduction Techniques based on Spectral Subtraction

Chapters 5, 6, and 7 described noise reduction techniques based on variations to the spectral subtraction method. These variations of the spectral subtraction algorithm (magnitude, power, etc.) were implemented in a generic form without the use of subframes and without synchronizing the process to the film rate. Informal listening tests indicated that these basic implementations were only effective at removing camera noise at rather high signal-to-noise ratios (>30 dB). At lower signal-to-noise ratios the artifacts

resulting from the processing made the systems unusable for the task of removing camera noise.

In all of the listening tests, the enhanced signal was compared directly with the original noisy input signal. Moreover, the difference of these two signals was also derived. This allows the listener to hear the portion of the original signal which was being removed by the spectral subtraction process. This ability was found to be particularly useful when trying to determine the best setting for a given parameter.

As described in Chapter 5, a very limiting artifact resulting from spectral subtraction is the musical noise. Track #10, which was derived using Boll's method, provides an example of musical noise. It should be noted that the quality of the speech signal is very good, but the musical noise makes the output signal unusable for a film soundtrack.

Track #10: Example of musical noise resulting from Boll's method.

Demo 6: Musical noise.

One way to overcome the problem of musical noise is to overestimate the level of the noise (see Section 5.4.5). However, if the overestimation is too high, the desired signal can become highly distorted. As a compromise, a minimum noise floor can be introduced to reduce the audibility of the musical noise without overly distorting the signal. Track #11 is the noisy speech segment (Track #1) processed by Boll's spectral subtraction algorithm with a minimum noise floor set to 30 dB below the level of the camera noise. In this example the audibility of the musical noise is greatly reduced compared to Track #10. There is however some distortion to the desired speech signal. Moreover, the camera noise is still audible and thus Track #11 is not usable in a film soundtrack.

Track #11: Spectral subtraction (Boll's method) with a minimum noise floor set to 30dB below the level of the camera noise.

Demo 7: Minimum noise floor.

In Section 5.4.5 a method for making the residual background noise more perceptually benign (i.e. hiss-like) was proposed. Track #12 is the noisy speech segment (Track #1) processed by Boll's spectral subtraction algorithm with a more benign residual noise. The minimum noise floor was set to 30 dB below the level of the camera noise. In this track, the structure of the camera noise has been greatly reduced and so the residual noise may be more acceptable.

Track #12: Spectral subtraction (Boll's method) with a perceptually benign noise floor set to 30dB below the level of the camera noise.

Demo 8: Benign minimum noise floor.

The various spectral subtraction algorithms were also implemented using two sub-frames. This provided a better estimate of the noise and also allowed for more aggressive processing to be applied using the overestimation parameter which was set independently for the two sub-frames. As a result, signals with poorer signal-to-noise ratios could be successfully processed using these systems. When the signal is also processed in sub-bands, the performance of the algorithm improves further. For example, a 4 subband system was implemented and tested. The highest 2 subbands were processed using 8 sub-frames. This allowed the processing to be very localized in these subbands and so the performance in the higher frequencies was noticeably improved. Specifically, the output signal did not suffer from as much high frequency roll-off. These findings were confirmed in the formal subjective tests.

The two perceptual models described in Chapter 7 (i.e., the high resolution Zwicker model and the Patterson-Moore model) were implemented and incorporated into the sub-band/sub-frame based spectral subtraction algorithm. It was found that both models provided a significant improvement to the noise reduction process as confirmed in the formal subjective tests. The effect of the perceptual model is to restore much of the high frequency speech signal that is removed when a high overestimation value is used. Moreover, the restored speech signal tends to have less temporal smearing when the perceptual model is used. These points will be discussed further in the next section.

To reduce the number of test items in the formal subjective test, a preliminary test was conducted to compare the performance of the two perceptual models (Tracks #13 to 18). Listeners compared the noisy speech segments processed by the spectral subtraction algorithms using the two models and were asked to indicate which segment they preferred. All listeners agreed that the Patterson-Moore model provided the better output. Low frequency noise components could be heard in the output of the spectral subtraction algorithm based on the high resolution Zwicker model. These noise components were more obvious when the input signal had a poorer signal-to-noise ratio. This is a reasonable finding since, as discussed in Chapter 7, the critical bands below 500 Hz are known to be too large and so the model would tend to overestimate the amount of masking available below 500 Hz. The unmasked noise resulting from the high resolution Zwicker

model is due in part to the very aggressive (noise reduction) processing used in this thesis. This aggressive processing is required in order to completely eliminate the camera noise as well as the musical noise. With less aggressive processing, this unmasking does not occur.

Track #13: *Ref* input signal (Track #1) processed using the high resolution Zwicker model.
Track #14: *Ref* input signal (Track #1) processed using the Patterson-Moore model.
Track #15: *Ref+6dB* input signal (Track #2) processed using the high resolution Zwicker model.
Track #16: *Ref+6dB* input signal (Track #2) processed using the Patterson-Moore model.
Track #17: *Ref+12dB* input signal (Track #3) processed using the high resolution Zwicker model.
Track #18: *Ref+12dB* input signal (Track #3) processed using the Patterson-Moore model.

Demo 9: Comparison of perceptual models.

To further compare the two perceptual models, it is instructive to listen to the masking thresholds predicted by the two models. To this end, a clean segment of speech was processed through the two models and the masking thresholds were determined. The masking thresholds were combined with the phase of the input signal to create audio signals which are representations of the masking thresholds due to the two models. Therefore, these signals represent the maximum amount of noise (as predicted by the two perceptual models) which can be masked by the input signal. Track #19 is the representation of the masking threshold for the high resolution Zwicker model and Track #20 is the representation of the masking threshold predicted by the Patterson-Moore model. The two masking thresholds sound quite different and the additional high and low frequency masking predicted by the high resolution Zwicker model can be heard in these tracks. This is an interesting result since the two perceptual models are intended to predict the same quantity (i.e., the masking threshold), yet their predictions are clearly different. This has important implications for other applications, such as perceptual audio codecs, which rely extensively on the predicted masking threshold. Most perceptual codecs are based on the critical band theory and thus the Zwicker model. The results described here strongly suggest that the Patterson-Moore model should be examined for these applications.

Track #19: Masking threshold predicted by the high resolution Zwicker model.
Track #20: Masking threshold predicted by the Patterson-Moore model.

Demo 10: Predicted masking thresholds.

By combining the various processes (subbands, sub-frames, perceptual model) described in this thesis, it is possible to obtain very good noise reduction on signals with relatively low signal-to-noise ratios. Therefore, the scheme developed in this thesis for reducing camera noise be deemed a success. This conclusion is confirmed by the results of the formal subjective tests described in the next section.

DAT recordings of noise from the IMAX™ camera were obtained and the algorithms developed in this thesis were applied to the recordings. It was found that the camera noise reduction system (based on subbands, sub-frames, perceptual model) was quite successful at removing the IMAX camera noise. However, due to the large size of the IMAX camera, its noise is generally much louder and so residual artifacts remained in the processed signal. One must decide whether to use more aggressive processing to completely eliminate the camera noise while audibly distorting the desired speech signal, or reduce the amount of processing thus allowing some residual camera noise to remain. Nonetheless, the methods proposed in this thesis appear to be very promising for reducing the IMAX camera noise.[‡]

Representatives at IMAX indicated that even though the noise reduction scheme cannot completely remove the camera noise without distorting the speech, it may still be a very useful tool in the automatic dialogue replacement process. Specifically, the noise reduction algorithm developed here could be used to significantly reduce (not eliminate) the camera noise prior to the ADR process. This would give the actors a much cleaner, and less distracting version of their dialogue to listen to during the ADR process.

8.3 Formal Subjective Test

A formal subjective test was conducted to evaluate the performance of the most promising camera noise reduction schemes to emerge from the thesis. Specifically, the subjective test was intended to evaluate the performance of a subband/sub-frame based spectral subtraction algorithm incorporating a perceptual model.

[‡] Unfortunately, due to copyright considerations, the demonstration CD does not include examples of the IMAX cameras.

The procedures and methods detailed in the ITU-R recommendation for subjective testing of audio systems with small impairments (ITU-R Rec. BS.1116 [179]) were followed. The subjective test methods outlined in BS.1116 are considered to be the most rigorous and sensitive for evaluating the quality of audio processing systems (i.e., algorithms), and are used extensively in the assessment of perceptual audio codecs. This recommendation addresses the performance of the playback system (amplifiers, loudspeakers, etc.), the acoustic characteristics of the listening environment (reverberation and background noise), assessment of listener expertise, the grading scale used to indicate subjective evaluations, and the methods of data analysis. The tests were carried out at the Audio Perception Lab of the Communications Research Centre in Ottawa, Canada which is one of the world's foremost subjective testing facilities. Further description of the use of BS.1116 as well as details regarding the Audio Perception Lab can be found in [180,181].

The subjective tests described in this chapter are perhaps the first such tests conducted to evaluate the quality of noise reduction algorithms. While it is true that other researchers have evaluated the quality of their noise reduction algorithms, they have not used the very sensitive methods used here. These sensitive methods are both warranted and necessary since the audio signals are full bandwidth signals (i.e., about 20 to 20000 Hz) and they will be used in an application (film soundtracks) which demands CD quality audio. Therefore, the use of telephony oriented subjective test methods, such as the Diagnostic Rhyme Test are not appropriate in this case.

A 15 s segment of male speech, consisting of four sentences from the Harvard Test Sequences [45], was used as the test material. The sentences were recorded in a quiet studio at the National Film Board of Canada in Montreal (see Section 3.5.6). The test segment was mixed with a recording of camera noise at 3 different levels (6 dB increments), thus providing 3 different signal-to-noise ratios. The 3 signal-to-noise ratios are referred to as *Ref*, *Ref+6dB*, and *Ref+12dB*, where *Ref* contains the highest level of camera noise. Relatively high levels of camera noise were used since it was felt that it would be easier for subjects to discriminate between algorithms. The levels of the noise were such that any noise reduction system which could successfully restore *Ref+12dB* should be capable of successfully eliminating camera noise in most applications.

These 3 test segments were then processed by 4 variants of the spectral subtraction algorithm, to produce a total of 12 test items for the subjective test. The 4 noise reduction algorithms consisted of: a standard spectral subtraction scheme based on Boll's method; a

subband/sub-frame spectral subtraction scheme based on Boll's method; a subband/sub-frame spectral subtraction scheme based on the Wiener filter; and a subband/sub-frame spectral subtraction scheme incorporating the Patterson-Moore based perceptual model described in Chapter 7. In processing the test segments for the subjective test, the over-estimation parameters were set such that all of the camera noise was eliminated as well as all of the musical noise. Since this is a necessary criteria for a successful camera noise reduction scheme, it was considered to be a suitable and fair criteria for determining the overestimation parameter.

The subjective test used the highly efficient within-subject (or repeated measures) design which is known to eliminate the effects of individual differences among subjects. A total of 21 subjects (17 male and 4 female) participated in the test and included many subjects who have previously shown a high level of expertise in audio subjective tests. Each subject conducted the test alone and the order in which each subject was exposed to the 12 test items was randomized thus eliminating the possibility of any time-related biases in the results.

A computer based playback system enabled the subject to instantaneously switch among any one of three versions of an auditory stimulus (see Figure 8.1) on each trial. Selecting button "A" on the screen produced the reference stimulus which was always known by the subject to be the clean speech segment. Clicking on button "B" or "C" produced either a hidden reference, identical to "A", or else a processed version of the test sequence. Which of "B" or "C" produced the hidden reference or the processed version was unknown and unpredictable to the subject from trial to trial.

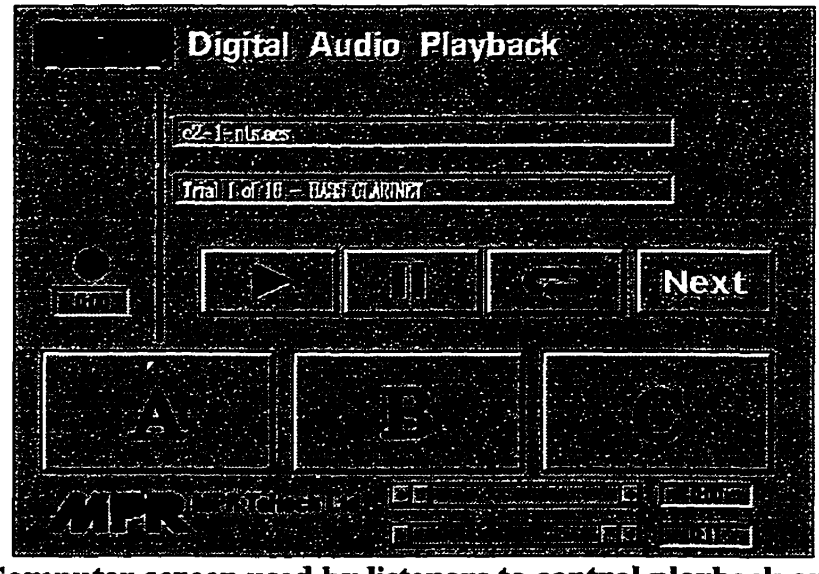


Figure 8.1 Computer screen used by listeners to control playback and switching.

The subject's task on each trial was to identify the processed version (on "B" or "C") and to grade its quality relative to that of the clean speech segment on "A". In the continuous grading scale used by the subjects, 1 to 1.9 represented an evaluation of varying degrees of a "very annoying" judgment, 2.0 to 2.9 covered the "annoying" range, 3.0 to 3.9 meant "slightly annoying", 4.0 to 4.9 was for judgments of "perceptible but not annoying", and 5.0 indicated "imperceptible". This is, in effect, a 41 grade continuous scale, with categorically labeled groupings for ease of orientation and to aid rating consistency throughout the experiment. The version judged to be the hidden reference (on "C" or "B") was given a grade of 5.0 ("imperceptible") so that on each trial one grade had to be "5.0".

During the blind-rating phase, each subject was free to take as much time as required on any trial, switching freely among the three stimuli as often as desired. The audio materials within a trial were time-synchronized so that the cross-fade when switching among "A", "B" and "C" was subjectively seamless. The subjects listened to the stimuli over loudspeakers since film soundtracks are typically auditioned in this manner.

Prior to conducting a formal double-blind listening test, each subject went through an extensive training session which allowed them to become familiar with the playback system, the experimental procedures, as well as the artifacts resulting from the noise reduction algorithms. The training process is outlined in BS.1116 and has been shown to provide a high degree of resolution and stability in subjective test results.

The test method described above using the A/B/C hidden reference double-blind approach allows the expertise of the subjects to be evaluated so that the results from non-expert subjects can be eliminated. A post-analysis showed that all of the subjects participating in this test demonstrated a high level of expertise and so the data from all subjects was included in the subsequent analysis.

An analysis-of-variance (ANOVA) was conducted on the test results and the main effects of the noise reduction algorithm, the signal-to-noise ratio, and the interaction between algorithms and signal-to-noise ratios were evaluated. The ANOVA indicated a highly significant effect ($p<0.001$) due to noise reduction algorithm and a highly significant effect ($p<0.001$) due to signal-to-noise ratio. The ANOVA also revealed that there was no significant interaction between the noise reduction algorithm and signal-to-noise ratio.

The results of the subjective test are shown in Figure 8.2. The horizontal axis indicates the 3 signal-to-noise ratios used in the test, while the vertical axis provides the MOS (mean opinion score) of the 21 subjects. Also included along the vertical axis are the 4 categorical descriptors used in the BS.1116 rating scale to describe the magnitude of any audible artifacts in the test items. Any two data points in the figure are statistically different ($p<0.05$) if their error bars do not overlap, while overlapping error bars indicate that the data points should be considered to be statistically identical.

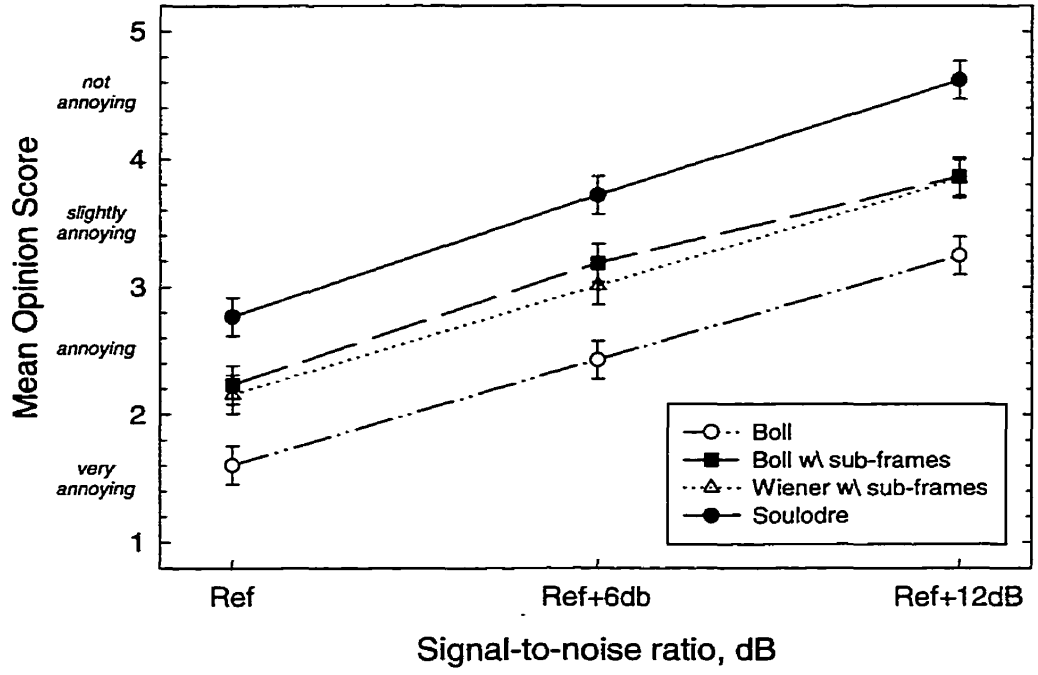


Figure 8.2 Results of subjective test.

The results of the subjective test provide a clear indication of the performance of the various noise reduction schemes. The lower curve represents the subjective performance of the traditional spectral subtraction algorithm based on Boll's method (i.e. magnitude subtraction). It can be seen that it provides the poorest subjective performance at each signal-to-noise ratio. The two middle curves represent the subband/sub-frame spectral subtraction algorithms based on Boll's method and the Wiener filter. It can be seen that these two algorithms provide a significant improvement in the quality of the resulting output signals. The two algorithms provide statistically identical results and so there does not appear to be any subjective benefit in using one algorithm instead of the other. The upper curve represents the spectral subtraction algorithm which incorporates the Patterson-Moore based perceptual model into the subband/sub-frame spectral subtraction algorithm (referred to in the figure as the Soulodre algorithm). The figure shows that this al-

gorithm provides a clear and consistent improvement in performance over the other noise reduction algorithms. Therefore, this algorithm provides the best overall performance for reducing camera noise.

There are several interesting details which can be gleaned from the figure. First, by dividing the spectral subtraction process into subbands and sub-frames, the improvement in the noise reduction algorithm is roughly equivalent to a 6dB increase in the signal-to-noise ratio of the input signal. That is, the performance of the subband/sub-frame spectral subtraction algorithms at a given signal-to-noise ratio is roughly equivalent to the performance of the traditional spectral subtraction algorithm at a 6dB higher signal-to-noise ratio. Similarly, the performance obtained by combining a perceptual model with the subband/sub-frame spectral subtraction algorithm (the Soulodre algorithm), is roughly equivalent to the performance of the traditional spectral subtraction algorithm operating at a 12 dB higher signal-to-noise ratio. These are significant gains in the signal-to-noise ratio since the performance of the spectral subtraction algorithms is very dependent on the initial signal-to-noise ratio of the input signal. For the *Ref+12dB* noise condition, this (Soulodre) algorithm had a MOS which fell in the *not annoying* range of the perceptual scale. In the evaluation of high quality audio systems, the ITU-R requires a score in this range in order for a system to meet "broadcast quality" (i.e. CD quality) requirements [182]. Essentially, a score in this range indicates that the system is providing an output which is virtually indistinguishable from the clean signal. Therefore, for signal-to-noise ratios at or above *Ref+12dB*, the noise reduction algorithm derived in this thesis satisfies this very stringent requirement. Given that *Ref+12dB* was intended to represent a relatively high level of camera noise, this result implies that this algorithm should be successful at removing camera noise under most (typical) conditions. Moreover, in many instances, music and sound effects will be mixed with the dialogue in the final soundtrack, thus masking any low-level residual artifacts.

It should also be noted that the improvement in the performance of the spectral subtraction algorithm which is obtained by including subbands, sub-frames, and a perceptual model is very robust. That is, the improvement in performance is consistent regardless of the initial signal-to-noise ratio of the input signal. This is confirmed by the fact that the ANOVA revealed that there was no significant interaction between the noise reduction algorithm and the signal-to-noise ratio of the input signal.

The audio sequences used in the subjective test are included on the demonstration CD. Track #21 is the segment of clean speech to which the camera noise was added.

This segment of speech was used as the reference (button "A") to which the subjects compared the various processed test items. Track #22 to Track #24 are the output of the spectral subtraction process for 3 levels of camera noise using the Boll method (i.e. magnitude subtraction) without subbands or sub-framing. Track #25 to Track #27 are the output of the spectral subtraction process using the Boll method with subbands and sub-framing. Track #28 to Track #30 are the output of the spectral subtraction process using the Wiener filter method with subbands and sub-framing. Track #31 to Track #33 are the output of the spectral subtraction process using both subbands and sub-framing as well as the Patterson-Moore based perceptual model (the Soulodre algorithm).

Track #21: Clean speech signal.
Track #22: *Ref* input signal (Track #1) processed using the Boll's method.
Track #23: *Ref+6dB* input signal (Track #2) processed using the Boll's method.
Track #24: *Ref+12dB* input signal (Track #3) processed using the Boll's method.
Track #25: *Ref* input signal (Track #1) processed using Boll's method with subbands/sub-frames.
Track #26: *Ref+6dB* input signal (Track #2) processed using Boll's method with subbands/sub-frames.
Track #27: *Ref+12dB* input signal (Track #3) processed using Boll's method with subbands/sub-frames.
Track #28: *Ref* input signal (Track #1) processed using Wiener filter method with subbands/sub-frames.
Track #29: *Ref+6dB* input signal (Track #2) processed using Wiener filter method with subbands/sub-frames.
Track #30: *Ref+12dB* input signal (Track #3) processed using Wiener filter method with subbands/sub-frames.
Track #31: *Ref* input signal (Track #3) processed using with subbands/sub-frames and the Patterson-Moore perceptual model.
Track #32: *Ref+6dB* input signal (Track #3) processed using with subbands/sub-frames and the Patterson-Moore perceptual model.
Track #33: *Ref+12dB* input signal (Track #3) processed using with subbands/sub-frames and the Patterson-Moore perceptual model.

Demo 11: Tracks used in formal subjective test.

Several types of artifacts can be clearly heard in the tracks listed above. The effects of temporal smearing become audible as the signal-to-noise ratio of the input signal becomes poorer and more aggressive processing is applied. Specifically, on Tracks #22, #25, and #28 certain syllables are quite smeared. Also, there is an obvious loss of high frequency information which becomes less severe as the signal-to-noise ratio of the input signal increases. Also, the loss of high frequencies is less severe for those tracks where

the perceptual model is included (i.e. Track #31 to Track #33). At lower signal-to-noise ratios, a *whistling* sound is superimposed on some portions of the speech.

It is interesting to note that some of the artifacts in the above tracks are more audible when auditioned over headphones rather than loudspeakers. For example, in Track #33, components of the noise are partially unmasked when auditioned over headphones, whereas these components are entirely masked in loudspeaker listening. This is a common observation in the subjective evaluation of perceptual-based audio codecs [179].

8.4 Conclusions

In this chapter, subjective evaluations were conducted of the various noise reduction schemes described in the thesis. A comparison of a spectral subtraction algorithm incorporating different perceptual models revealed that the Patterson-Moore based model tended to perform better than the Zwicker based model for the noise reduction application. Audio representations of the masking thresholds predicted by the two models revealed clear and important differences. This result has significant implications for other applications (e.g. perceptual audio codecs) not related to noise reduction.

A formal subjective test of the most promising schemes demonstrated that a sub-band/sub-frame based spectral subtraction algorithm incorporating the Patterson-Moore based perceptual model provides the best noise reduction performance for camera noise. This algorithm performs as well as a traditional spectral subtraction algorithm operating on an input signal having a 12 dB higher signal-to-noise ratio. It was also shown that the improved performance (over traditional spectral subtraction) due to this algorithm is robust and is relatively independent of the initial signal-to-noise ratio of the input signal. The results of the formal subjective test support the philosophy taken in this thesis of minimizing the amount of processing applied to the signal.

The results of the formal subjective test indicate that the newly developed algorithm should be successful at removing camera noise under typical conditions without audibly distorting the signal. Therefore, this algorithm satisfies the requirements for a successful camera noise reduction system.

9. CONCLUDING REMARKS

In this thesis, the problem of camera noise corrupting film soundtracks was investigated and an effective method for reducing the noise was developed. Section 9.1 summarizes the key points of the thesis, while Section 9.2 looks at future research directions.

9.1 Summary

The problem of camera noise in film soundtracks was introduced and the requirements for a noise reduction scheme were outlined. A successful scheme must render the camera noise inaudible while not significantly distorting the underlying speech signal. Another requirement is that it must be a single-input system which can be applied in post-production. This requirement adds a significant degree of difficulty to the task.

In Chapter 2, the methods currently used to limit or eliminate camera noise in film soundtracks were described. Methods based on microphone techniques attempt to focus-in on the desired signal either by using a highly directional microphone or by placing the microphone as close as possible to the actors. In another approach, the camera noise is limited at the source by placing the camera behind an acoustic barrier. An analogue signal processing device (the Dolby 430 Series) was described which is sometimes used to try to reduce camera noise. The device was not designed with the intent of reducing camera noise, and so its usefulness in this application is limited.

A comprehensive characterization of camera noise was described in Chapter 3. An examination of the time waveform revealed that the camera noise consists of a series of pulses coinciding with the film rate of the camera (24 frames per second). Each pulse consists of an initial peak followed by an interval of noise. While the pulses showed similarities, it was seen that the pulses are in fact different. Directivity measurements were also made, and it was found that the power spectrum of the camera noise changes with the angle of the measurement.

Several factors were evaluated to determine whether they caused any variation in the camera noise. It was found that the type of lens mounted on the camera had a small effect, whereas the film stock caused much greater changes to the noise. Measurements taken over time demonstrated that the power spectrum of the camera noise did not change significantly within a given reel of film. Measurements were also conducted on three IMAX cameras. These cameras were very different (physically, and in their intended ap-

plication) from each other and were also very different from the NFB camera. Yet each camera exhibited similar characteristics which can be exploited in the noise reduction process.

It was shown that, due to its physical size, as well as its many mechanical components, the camera behaves as a distributed noise source. This finding has important implications regarding the possible success of certain noise reduction schemes. The IMAX cameras also behaved as distributed noise sources. The consistency in the characteristics of the cameras measured in the this chapter support the notion that the noise reduction methods developed in the thesis should be widely applicable to other types of cameras.

A mathematical model was developed to describe camera noise. The model divides the camera noise into two parts: a periodic component, and a cyclical random component. This allowed the noise reduction schemes to be described in terms of their ability to reduce either or both of the components.

In Chapter 4, ANC based methods for reducing camera noise were investigated and a review of the theory behind the LMS algorithm and some of its variants was provided. The potential limitations (extraneous noise sources, rate of convergence, leakage of desired signal into reference input, etc.) of the adaptive noise cancellation method were identified. It was shown that the performance of the ANC method is dependent on the coherence between the two input signals. The coherence is reduced in the presence of a diffuse sound field and it was shown that a distributed noise source acts as a diffuse noise field. Since the camera is a distributed noise source, the inputs to an ANC system (for reducing camera noise) have low coherence. Therefore, it can be concluded that, for the purpose of reducing camera noise, an ANC system will have only limited success. This conclusion was confirmed through experimental results.

Blind signal separation methods were reviewed and their relation to ANC was highlighted. The effect of microphone spacing on the performance of a blind signal separation system based on second order statistics was considered. It was shown that blind signal separation is also dependent on the coherence between the input signals, and therefore its performance will be limited by the distributed nature of camera noise.

To satisfy the single-input requirement for the camera noise reduction system, an adaptive noise canceling scheme using a synthesized reference input was proposed. The method relies on the noise having a high inter-pulse correlation and is intended to reduce the level of the periodic component of the camera noise. Unfortunately, an analysis of the

camera noise found rather low inter-pulse correlation and thus the approach is largely unsuccessful for the present application. The poor performance of this method was shown to be due in part, to the jitter in the times of arrival of the individual camera noise pulses. To resolve this matter, a method for synchronizing the ANC process to the camera noise was devised. This method significantly increased the inter-pulse correlation (primarily in the higher frequencies) and consistently provided between 10 and 15 dB of noise reduction.

In Chapter 5, signal enhancement techniques based on estimating the short-time spectral magnitude of the signal were investigated. The mathematical foundation for the spectral subtraction process was derived for the method proposed by Boll. Spectral subtraction reduces both the periodic and cyclical random components of camera noise. It was shown that spectral subtraction is equivalent to a zero-phase filter and that this interpretation allows for both a better understanding and a generalization of the process.

The limitations of the method were identified and the artifacts resulting from spectral subtraction were described. Several new modifications to the traditional spectral subtraction algorithm were proposed to minimize the audibility of these artifacts. The minimum spectral floor proposed by Berouti *et al.* for reducing the audibility of musical noise was extended to make the noise floor more perceptually benign. A modified version of the survival algorithm devised by Vaseghi and Frayling-Cork was proposed which was found to provide better reduction of musical noise. An overestimation parameter based on the mean and variance of the noise was proposed, as was a means of reducing time aliasing effects caused by modifications to the spectrum of a signal. An analysis/synthesis windowing function (when performing the FFT/IFFT operations) was added to remove the discontinuities at the boundaries of overlapping processing frames. Use of these extensions is not limited to the camera noise problem, and can improve the performance of the spectral subtraction algorithm in general.

In Chapter 6 general mathematical framework for integrating subbands and subframes into the spectral subtraction algorithm was derived based on the use of quadrature mirror filter banks. It was shown that matching the noise reduction process (in the time-frequency plane) to the noise can significantly improve the performance of the spectral subtraction algorithm by reducing all forms of audible artifacts. By directing the processing to those portions of the noise which require it most, the overall amount of processing applied to a signal can be significantly reduced. This approach was generalized using non-equal sub-frames. The need for window alignment and frame synchronization

when using sub-frames was demonstrated and a simple means of maintaining frame synchronization was proposed. This general philosophy of minimizing the amount of processing applied to the signal in a given subband/sub-frame is carried throughout the thesis.

In Chapter 7 the addition of a perceptual model to the spectral subtraction process was investigated. The perceptual model was used to determine which portions of the noise are audible and which are being masked by the desired signal. The noise reduction process is then limited to reducing those portions of the noise which are audible. As such, the overall amount of processing applied to the signal is minimized from a perceptual point of view and thus the levels of the artifacts are also reduced.

Certain limitations of the critical band based model used by Tsoukalas *et al.* were identified and two other models of simultaneous masking were considered. The high resolution Zwicker model and the Patterson-Moore model were compared from a mathematical point of view by mapping them from their basilar membrane representations to the linear frequency domain. The models were shown to produce significantly different estimates of the auditory masking threshold.

A new perceptual model was developed based on the Patterson-Moore model for simultaneous masking. The model also contained several new components derived in this thesis to account for the filtering effects of the outer and middle ear, the internal noise floor of the internal ear, forward masking as a function of level and frequency, nonlinear addition of masking, and the interaction between the auditory filters and the window used in the Fourier transform. The new model was incorporated into a subband/sub-frame based spectral subtraction algorithm and a new method for estimating the clean signal was proposed. This implementation of spectral subtraction with the perceptual model was found to provide superior performance over traditional implementations.

Chapter 8 provided a subjective evaluation of the performance of the various noise reduction algorithms derived in the thesis. Listening tests confirm the differences between the auditory masking thresholds predicted by the two perceptual models. The spectral subtraction algorithm operating with the Patterson-Moore based perceptual model was found to perform better than the algorithm based on the Zwicker model. A formal subjective test using the most rigorous and sensitive methods was conducted to evaluate the more promising noise reduction algorithms. The Boll method and the Wiener filter method of spectral subtraction were found to be subjectively equivalent in their performance. The results of the formal subjective test clearly show that the subband/sub-frame based spectral subtraction algorithm with the new Patterson-Moore based percep-

tual model provides the best performance overall. The results demonstrate a significant improvement in the performance of the spectral subtraction algorithm due to the use of subbands and sub-frames, as well as the use of a perceptual model. It was therefore concluded that the methods developed in the thesis meet the requirements for a successful camera noise reduction system.

This thesis appears to constitute the first formal effort to apply adaptive signal processing techniques to the problem of reducing camera noise in film soundtracks. In the thesis, the characteristics of camera noise were thoroughly analyzed and many different approaches to reducing camera noise were investigated. The method based on spectral subtraction using subbands, sub-frames and a new perceptual model was shown to provide very good performance. The application of adaptive signal processing techniques to the problem of camera noise is in its infancy, and one might compare its current state to that of gramophone restoration about 10 to 15 year ago. It is likely that, as has occurred with gramophone restoration, other researchers will investigate the camera noise problem and develop new methods for improving the quality of the noise reduction algorithms. Hopefully, the work in this thesis will provide some insight to these researchers about which methods are most likely to be successful.

The work in this thesis related to the analysis and development of a new perceptual model is of potential benefit to many applications not related to camera noise. The model can be incorporated into any spectral subtraction algorithm for reducing a noise which is corrupting a speech or music signal. Moreover, since the model uses an auditory filter approach, it is relatively straightforward to develop customized implementations of the algorithm which might be useful in some applications. The new perceptual model also offers interesting new possibilities in the fields of perceptual audio codecs as well as perceptual-based objective measurement systems.

9.2 Future Research Directions

There are several issues which should be addressed in order to improve the performance of the signal enhancement system developed in this thesis. In this section, some of the more important issues are identified.

9.2.1 Realtime Implementation

The work in this thesis provided a generalized form of the spectral subtraction process as well as several extensions intended to minimize specific artifacts created by the spectral

subtraction process. Moreover, a subband/sub-frame framework was introduced as was a perceptual model. As a result, there are numerous parameters which must be adjusted in order to obtain the highest degree of noise suppression without causing significant distortion to the underlying speech signal. In this thesis, all processing was done on a non-realtime basis, thus making it rather difficult to fine tune the values of these parameters. Furthermore, it was not practical to process large amounts of data.

At this stage, a very useful next step in the research would be to develop a realtime implementation of the process. This would allow the effects of each parameter in the process to be evaluated quickly and efficiently. Furthermore, this is the most sensible means of fully determining the required tradeoffs between the various parameters.

The computational requirements for most parts of the noise reduction scheme devised in this thesis are not excessively demanding. The most demanding portions of the process are those related to the calculation of the auditory masking threshold required for each block of data. As implemented in the thesis, this requires the multiplication of a 1 by N vector with an N by N matrix. Present day DSP chips include specific instructions designed to rapidly perform these operations. Furthermore, significant computational savings could be had by calculating the auditory masking threshold in the basilar membrane domain where the (pitch) resolution is lower and so the number of points in the calculations is much less than N. It seems probable that, given the current state of the art, the entire noise reduction process could be performed in realtime on a single high-end DSP processor. Due to requirements for high quality audio output (i.e. 16 bits), the processor should have an internal resolution of no less than 24 bits.

9.2.2 Improved Perceptual Model

The use of a perceptual model was found to provide a significant improvement in the performance of the spectral subtraction algorithm. Moreover, the new model based on the Patterson-Moore simultaneous masking model was found to provide better performance than the model based on the Zwicker model. It should be noted that these models were not developed with engineering applications in mind, and so from an engineering point of view, these two models leave many important questions unanswered. It would be very useful to perform a series of fundamental experiments related to auditory masking which are designed with engineering applications in mind. Such experiments would include the effects of frame-based processing, transform-based processing, windowing effects, etc.

For example, work which is currently being done in the area of very low bitrate perceptual audio codecs may be useful in this regard. At very low bit rates it is inevitable that the effects of quantization (e.g. quantization noise, temporal smearing, timbre effects, etc.) will be audible. Research is currently being done to investigate strategies for minimizing the perceived annoyance due these inevitable artifacts. The results of these studies may help designers of noise reduction algorithms to find the best balance between residual noise and artifacts in the signal. The author is currently conducting such experiments, but there are many questions to be answered.

9.2.3 Improved Reduction of Periodic Noise Component

In Chapter 4 a method was developed for reducing the level of the periodic component of camera noise. This method was shown to be inadequate by itself as a means of reducing camera noise. It seems sensible to use this method as a pre-processor to the spectral subtraction based method. However, this approach did not perform well because the ANC based method developed in Chapter 4 does not adequately reduce the periodic noise component during intervals of speech activity. It is possible that the removal of the periodic component of the camera noise could be improved at lower frequencies (especially during speech activity) by using the method proposed by Godsill and Tan [42] for removing low frequency transient noise from gramophone recordings. They model the transient noise as an autoregressive process and use a Kalman filter approach to remove the noise. If this method can provide a more consistent reduction of the periodic component, including during speech activity, then it may be useful as a pre-processor to a spectral subtraction algorithm. This could be quite beneficial in situations where the level of the camera noise is excessively high. It is not clear how their method would perform at higher frequencies. Nonetheless, their method should be investigated.

9.2.4 Use of Discrete Cosine Transform

In their work on noise reduction using a signal subspace approach, Ephraim and Van Trees used the KLT to decompose the noisy vector into two subspaces. This method was reported to provide an enhanced signal which is free from musical noise. Use of the KLT however, is computationally demanding and does not lend itself easily to integration with a perceptual model. As a possible compromise, it would be worthwhile to investigate the use of a discrete cosine transform in the signal subspace approach. The DCT is known to give a good approximation to the KLT yet it is computationally efficient. Moreover, the DCT lends itself more readily to integration with a perceptual model.

9.2.5 Increased Number of Frequency Subbands

The use of a combination of sub-frames and subband filtering was seen to provide a significant improvement in the performance of the spectral subtraction system. In the simulations performed in this thesis, four frequency subbands were used. The use of more subbands at lower frequencies (below 6 kHz) should be investigated as a possible means of achieving better performance at these frequencies. This may provide additional improvement when used in conjunction with a perceptual model since upward masking is the dominant form of masking and so it is important to have the best possible estimate of the signal at low frequencies.

9.2.6 Phase Estimation

Vary showed the relation between the maximum expected deviation in phase and the signal-to-noise ratio of the input signal. His results suggest a possible means of trying to estimate the short-time phase of the desired signal. Vary's work provides bounds which would help in the estimation of the phase. It is expected that a more accurate estimate of the phase could provide improved performance (particularly at low signal-to-noise ratios) by reducing the artifacts related to temporal aliasing.

9.2.7 Multiple Passes of the Spectral Subtraction Process

While investigating the effects of the various parameters in the spectral subtraction process, it was found that if only a moderate amount of noise suppression was applied to the corrupted signal, no musical noise resulted. The amount of noise suppression was controlled by the parameters α, β, and γ as shown in Figures 5.3 to 5.6. That is, α, β, and γ were set so that the slope of the noise suppression curve did not become too steep thus causing musical noise. Of course, the remaining camera noise was still at an unacceptable level.

To obtain more noise suppression, the processed signal was re-processed through the spectral subtraction system. A new estimate of the background noise was derived from the residual noise at the output of the first spectral subtraction process. This procedure was repeated over several iterations.

Although this approach was not entirely successful, it did show enough promise to warrant further investigation. Since musical noise results from the non-linearity of the spectral subtraction process, it seems reasonable that it could be reduced by decreasing the severity of the non-linearity.

9.2.8 Higher Order Statistics

Though not reported in the thesis, the use of higher order statistical methods was investigated as a possible means of taking advantage of the repetitive (cyclical) nature of camera noise. Higher order statistics are frequently used when dealing with cyclostationary signals, and therefore it was felt that they could be useful in reducing camera noise [183,184,185,186,187,188,189,190,191]. A major drawback with signal processing methods based on higher order statistics (such as the bispectrum or trispectrum) is the computational requirements involved. Calculation of these statistical measures is far more demanding than second-order quantities. Also, they must be calculated over many cycles in order to obtain accurate measurements. Moreover, since the rate of repetition of camera noise is relatively slow (24 frames per second), one must examine the signal over a large number of samples in order to include a sufficient number of "cycles" for the noise to appear cyclical. This implies that a camera noise reduction scheme based on higher order statistics would have to deal with very large quantities of data and the processing of this data would be very computationally demanding. Given the author's current computing capabilities, it was not possible to implement algorithms based on the use of higher order statistics. As greater computing power (and memory) becomes available, it may be worthwhile to investigate these methods.

9.3 Epilog

In this dissertation a signal processing method was proposed which removes camera noise from film soundtracks. The method uses a technique known as spectral subtraction which is based on estimating the short-time spectral magnitude of the desired signal. The basic spectral subtraction process however, creates audible artifacts which are often more disturbing than the original noise and thus new algorithms were proposed to minimize these artifacts. The spectral subtraction process was also extended to take advantage of the cyclical or repetitive nature of camera noise. Sub-frames, which were synchronized and aligned to the interfering noise, used in conjunction with frequency subbands were found to significantly improve the noise reduction process. The use of subbands and sub-frames permits the noise reduction process to be better matched to the noise in the time-frequency plane. This in turn allows the overall amount of processing applied to the signal to be reduced, thus reducing the resulting artifacts. The noise reduction process was further improved by including a perceptual model which allows further reduction in the amount of processing applied to the signal. A subband/sub-frame based spectral subtraction algorithm using a perceptual model provided a means of successfully removing cam-

era noise from film soundtracks without adversely affecting the quality of the underlying signal. In a formal subjective test the proposed method was shown to work well even in the presence of relatively high levels of camera noise.

While the work in this thesis examined the specific problem of reducing camera noise, the results could be extended to other applications which require a signal to be enhanced in the presence of a repetitive noise. Also, much of the work can be applied to the general problem of noise reduction in audio signals, while other aspects of the thesis are directly useful to other applications.

APPENDIX A

The power spectrum of the camera noise can be averaged over short intervals of time related to the pulse rate of the camera (i.e. 2000 samples), or it can be viewed over longer intervals spanning many pulses. The choice of whether to view the power spectrum of the interfering noise over a shorter or longer time interval depends on the application. For example, some noise reduction schemes, such as spectral subtraction described in Chapter 5, process the signal over short intervals wherein the signal (typically speech) is considered to be stationary.

It was seen in Figure 3.7 that, when viewed over a time interval equal to the period of the noise pulses, there was no obvious structure to the camera noise. Specifically, there were no spectral lines that would indicate a harmonic structure to the noise. However, given that the camera operates at a rate of 24 frames per second, it is reasonable to assume that a power spectrum measured over several noise pulses would reveal spectral lines spaced 24 Hz apart. Indeed, these spectral lines do occur, but only if the power spectrum is measured over many pulses of the camera noise.

Figure A.1 shows the power spectrum of the camera noise measured over 16 pulses of the camera noise (i.e. 2/3rds of a second). Spectral lines spaced 24 Hz apart can be seen, although the amplitudes of the spectral lines do not appear to follow a constant pattern. The amplitudes of the spectral lines tend to decrease with increasing frequency. Although not seen in the figure, the spectral lines tend to disappear above about 2500 Hz, even though the spectrogram of Figure 3.8 showed significant noise energy above this frequency associated with the onsets of the pulses. Furthermore, the magnitudes of the spectral lines vary with the angle of the measurement, the film stock, and the type of lens mounted on the camera.

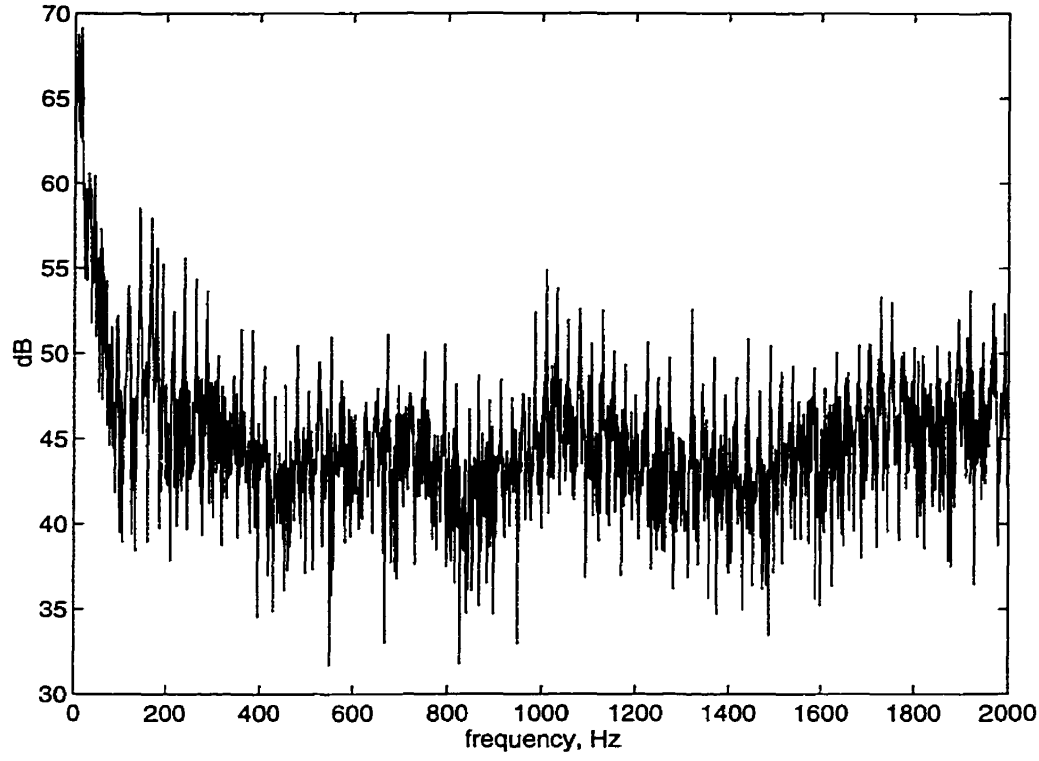


Figure A.1 Typical power spectrum of the camera noise measured over 16 noise pulses.

There is an underlying broadband noise that can be seen between the spectral lines. This is due to the fact that the camera noise is effectively a series of noise bursts and that each burst is unique. Furthermore, the camera noise is made up of other components that are not directly related to the 24 frames per second film rate.

The power spectrum of the camera noise can look quite different depending on the length of the time interval over which it is measured. For the spectral subtraction scheme used in this thesis it is necessary to process the signal over short time intervals, and therefore, the power spectrum must also be measured over these short intervals.

APPENDIX B

Twenty sentences from the 1965 revised list of phonetically balanced sentences (Harvard Sentences [45]) were used in evaluating the performance of the various noise reduction schemes described in this thesis. The Harvard Sentences consist of 720 phonetically balanced sentences which are grouped into 72 lists each containing 10 sentences. List 1 and list 5 were used in the tests, and were recorded as described in Section 3.5.6. These 20 sentences are listed below.

List 1:

1. The birch canoe slid on the smooth planks.
2. Glue the sheet to the dark blue background.
3. It's easy to tell the depth of a well.
4. These days a chicken leg is a rare dish.
5. Rice is often served in round bowls.
6. The juice of lemons makes fine punch.
7. The box was thrown beside the parked truck.
8. The hogs were fed chopped corn and garbage.
9. Four hours of steady work faced us.
10. A large size in stockings is hard to sell.

List 5:

1. A king ruled the state in the early days.
2. The ship was torn apart on the sharp reef.
3. Sickness kept him home the third week.
4. The wide road shimmered in the hot sun.
5. The lazy cow lay in the cool grass.
6. Lift the square stone over the fence.
7. The rope will bind the seven books at once.
8. Hop over the fence and plunge in.
9. The friendly gang left the drug store.
10. Mesh wire keeps chicks inside.

REFERENCES

[1] A. Nisbett, *The Sound Studio*. Oxford: Focal Press, 1993, 5th ed.

[2] J. M. Weaver, "Master re-recording mixer — Richard Portman," Mix, vol. 19, no. 9, pp. 20–30, Sept. 1995.

[3] D. Simpson, The National Film Board of Canada, *Private Communication.*

[4] J. P. Vialard, The National Film Board of Canada, *Private Communication.*

[5] G. Harris, IMAX™ Corporation, *Private Communication.*

[6] E. M. Di Giulio, E. C. Manderfeld, and G. A. Mitchell, "An historical survey of the professional motion-picture camera," *J. SMPTE*, vol. 85, pp. 487–492, July 1967.

[7] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeildler, E. Dong, Jr., and R. C. Goodlin, "Adaptive noise canceling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.

[8] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs: Prentice-Hall, 1986.

[9] P. M. Clarkson, *Optimal and Adaptive Signal Processing*. Boca Raton: CRC Press, 1993.

[10] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Englewood Cliffs: Prentice–Hall, 1985.

[11] O. Macchi, *Adaptive Processing: The Least Mean Squares Approach with Applications in Transmission*, Chichester: John Wiley & Sons, 1995.

[12] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Processing speech signals to attenuate interference," *IEEE Symp. Speech Recognition*, pp. 292–293, Apr. 1974.

[13] S. F. Boll, "Suppression of noise in speech using the SABER method," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 606–609, Apr. 1978.

[14] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[15] R. J. McAulay and M. L. Malpass, "A real-time noise suppression filter for speech enhancement and robust channel vocoding," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 699–702, Apr. 1980.

[16] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 5, pp. 471–472, Oct. 1978.

[17] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 208–211, Apr. 1979.

[18] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.

[19] P. Vary, "Noise suppression by spectral magnitude estimation — mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387–400, July 1985.

[20] S. V. Vaseghi and R. Frayling-Cork, "Restoration of old gramophone recordings," *J. Audio Eng. Soc.*, vol. 40, no. 10, pp. 791–801, Oct. 1992.

[21] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

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

[23] D. E Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 5, no. 6, pp. 497–514, Nov. 1997.

[24] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov models and the projection, for robust speech recognition in cars," *Speech Communications*, vol. 11, no. 2/3, pp. 215–228, June 1992.

[25] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, vol. 2, pp. 363–366, Apr. 1993.

[26] R. Le Bouquin, "Enhancement of noisy speech signals: Application to mobile communications," *Speech Communications*, vol. 18, pp. 3–19, Jan. 1996.

[27] R. D. Preuss, "A frequency domain noise cancellation preprocessor for narrowband speech communications systems," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 212–215, Apr. 1979.

[28] S. Maitra, "Reducing the effect of background noise for low-bit-rate voice digitizers," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 696–698, Apr. 1980.

[29] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[30] O. Cappé, "Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 1, pp. 84–93, Jan. 1995.

[31] W. C. Sabine, *Collected Papers on Acoustics*, edited by T. Lyman, New York: Dover, 1964.

[32] G. A. Soulodre and J. S. Bradley, "Subjective evaluation of new room acoustic measures," *J. Acoust. Soc. Amer.*, vol. 98, no. 1, pp 294–301, July 1995.

[33] G. W. Elko, "Directional microphones for teleconferencing," *J. Acoust. Soc. Amer.*, vol. 101, No. 5, Pt. 2, p. 3118 (A), May 1997.

[34] L. L. Beranek, *Acoustics*, New York: McGraw Hill, 1954.

[35] Dolby 430 Series Background Noise Suppressor System - User's Manual.

[36] John M. Woram, *Sound Recording Handbook*, Indianapolis: Howard W. Sams & Co., 1989.

[37] Tim Haupt, Mind's Eye Productions, *Private Communication.*

[38] S. V. Vaseghi and P. J. W. Rayner, "A new application of adaptive filters for restoration of archived gramaphone recordings", in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 2548–2551, vol. 5, Apr. 1988.

[39] S. J. Godsill, *The Restoration of Degraded Audio Signals*, Ph.D. Thesis, University of Cambridge, U.K., 1993.

[40] S. J. Godsill and P. J. W. Rayner, "A bayesian approach to the restoration of degraded audio signals", *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4. pp. 267–278, July 1995.

[41] S. J. Godsill and P. J. W. Rayner, "Robust noise reduction for speech and audio signals", in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 625–628, Apr. 1996.

[42] S. J. Godsill and C. H. Tan, "Removal of low frequency transient noise from old recordings using model-based signal separation techniques", in *Proc. IEEE workshop on applications of signal processing to audio and acoustics*, Oct. 1997.

[43] B. Lambert, The Walt Disney Company, *Private Communication*, 1996.

[44] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms", *IEEE Transactions on Audio and Electroacoustics*, AU-15, no. 2, pp. 70–73, June 1967.

[45] IEEE Standards Publication No. 297, IEEE Recommended Practice for Speech Quality Measurements, *IEEE Trans. on Audio and Electroacoustics*, vol. AU-17, no. 3, Sept. 1969.

[46] A. G. Piersol, "Use of coherence and phase data between two receivers in evaluation of noise environments," *J. Sound Vib.* vol. 56, no. 2, pp. 215–228, Jan. 1978.

[47] B. Widrow and M. Hoff, Jr., "Adaptive switching circuits," in *IRE WESCON Conv. Rec.*, Pt. 4, pp. 69–104, 1960.

[48] S. M. Kuo and D. R. Morgan, *Active Noise Control Systems*, New York: John Wiley & Sons, 1996.

[49] M. Bellanger, *Adaptive Filters and Signal Analysis*, New York: Dekker, 1988.

[50] D. S. Watkins, *Fundamentals of Matrix Computations*, New York: John Wiley & Sons, 1991.

[51] M. Dentino, J. McCool, and B. Widrow, "Adaptive filtering in the frequency domain," *Proc. IEEE*, vol. 66, pp. 1658–1659, Dec. 1978.

[52] S. S. Narayan, A. M. Peterson, and M. J. Narasimha, "Transform domain LMS algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, no. 3, pp. 609–615, June 1983.

[53] J. C. Lee and C. K. Un, "Performance of transform-domain LMS adaptive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 3, pp. 499–510, June 1986.

[54] D. F. Marshall, W. K. Jenkins, and J. J. Murphy, "The use of orthogonal transforms for improving performance of adaptive filters," *IEEE Trans. Circuits and Systems*, vol. 36, no. 4, pp. 474–484, Apr. 1989.

[55] A. Gersho, "Adaptive equalization of highly dispersive channels for data transmission," *Bell Syst. Tech. J.*, vol. 48, pp. 55–70, Jan. 1969.

[56] F. Beaufays, "Transform-domain adaptive filters: an analytical approach", *IEEE Trans. Signal Processing*, vol. 43, no. 2, pp. 422–431, Feb. 1995.

[57] M. Doroslovacki and H. Fan, "Wavelet-based adaptive filtering," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 488–491, vol. 3, Apr. 1993.

[58] N. Erdol and F. Basbug, "Performance of wavelet transform based adaptive filters," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 500–503, vol. 3, Apr. 1993.

[59] S. Hosur and A. H. Tewfik, "Wavelet transform domain LMS algorithm," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 508–510, vol. 3, Apr. 1993.

[60] S. Hosur and A. H. Tewfik, "Wavelet transform domain adaptive FIR filtering," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 617–630, March 1997.

[61] S. Attallah and M. Najim, "On the convergence enhancement of the wavelet transform based LMS," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 973–976, Apr. 1995.

[62] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, pp. 14–37, Jan. 1992.

[63] S. F. Boll and D. C. Pulsipher, "Suppression of acoustic noise in speech using two microphone adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 6, pp. 752–753, Dec. 1980.

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

[65] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[66] D. Pulsipher, S. F. Boll, C. Rushforth, and L. Timothy, "Reduction of nonstationary acoustic noise in speech using LMS adaptive noise cancelling," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 204–207, Apr. 1979.

[67] W. A. Harrison, J. S. Lim, and E. Singer, "Adaptive noise cancellation in a fighter cockpit environment," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 18A.4.1–18A.4.4, March 1984.

[68] W. A. Harrison, J. S. Lim, and E. Singer, "A new application of adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 1, pp. 21–27, Feb. 1986.

[69] P. Darlington, P. D. Wheeler, and G. A. Powell, "Adaptive noise reduction in aircraft communication systems," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 19.2.1–19.2.4, March 1985.

[70] G. A. Powell, P. Darlington, and P. D. Wheeler, "Practical adaptive noise reduction in the aircraft cockpit environment," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 6.2.1–6.2.4, Apr. 1987.

[71] J. J. Rodriguez, J. S. Lim, and E. Singer, "Adaptive noise reduction in aircraft communications systems," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 6.1.1–6.1.4, Apr. 1987

[72] G. Elko, "Adaptive noise cancellation with directional microphones," in *Proc. IEEE workshop on applications of signal processing to audio and acoustics*, Oct. 1997.

[73] G. Elko, Lucent Technologies, *Private Communication*.

[74] J. Blauert, *Spatial Hearing*, Cambridge: MIT Press, 1983.

[75] C. Jutten and J. Herault, "Blind separation of sources, Part I: an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, July 1991.

[76] T. Nomura, M. Eguchi, H. Niwamoto, H. Kokubo, and M. Miyamoto, "An extension of the Herault-Jutten network to signals including delays for blind separation," *IEEE Workshop on Neural Networks for Signal Processing*, pp. 443–452, Sept. 1996.

[77] A. Cichocki, R. E. Bogner, L. Moszczynski, and K. Pope, "Modified Herault-Jutten algorithms for blind separation of sources," *Digital Signal Processing*, vol. 7, no. 2, pp. 80–93, Apr. 1997.

[78] J.-F. Cardoso, "Source separation using higher order statistics," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 2109–2112, May 1989.

[79] H-L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Processing*, vol. 45, no. 2, pp. 209–229, Aug. 1995.

[80] S. Shamsunder and G. B. Giannakis, "Multichannel blind signal separation and reconstruction," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 6. pp. 515–528, Nov. 1997.

[81] A. J. Bell and T. J. Sejnowski, "Blind separation and blind deconvolution: an information-theoretic approach," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 3415–3418, May 1995.

[82] K. Torkkola, "Blind separation of convolved sources based on information maximization," *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 423–432, Sept. 1996.

[83] K. Torkkola, "Blind separation of delayed sources based on information maximization," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, May 1996.

[84] T.-W. Lee, A. J. Bell, and R. H. Lambert, "Blind separation of delayed and convolved sources," *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA., 1997.

[85] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *Proc. IEEE workshop on applications of signal processing to audio and acoustics*, Oct. 1997.

[86] C. N. Canagarajah, *Digital Signal Processing Technique for Speech Enhancement in Hearing Aids*, Ph.D. thesis, Cambridge University, U.K., 1993.

[87] D. Van Compernolle and S. Van Gerven, "On the use of decorrelation in scalar signal separation," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 57–60, Apr. 1994.

[88] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation", *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 4. pp. 405–413, Oct. 1993.

[89] D. C. B. Chan, *Blind Signal Separation*, Ph.D. thesis, University of Cambridge, U.K., 1997.

[90] D. Yellin and E. Weinstein, "Multichannel signal separation: methods and analysis", *IEEE Trans. Signal Processing*, vol. 44, no. 1, pp. 106–118, Jan. 1996.

[91] D. C. B. Chan, P. J. W. Rayner, and S. J. Godsill, "Multi-channel blind signal separation by decorrelation", in *Proc. IEEE workshop on applications of signal processing to audio and acoustics*, Oct. 1995.

[92] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations", *Phys. Rev. Lett.*, vol. 72, no. 23, pp. 3634–3637, June 1994.

[93] F. Ehlers and H. G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment", *IEEE Trans. Signal Processing*. 45, no. 10, Oct. 1997.

[94] D. C. B. Chan, P. J. W. Rayner, and S. J. Godsill, "Multi-channel signal separation", in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 649–652, May 1996.

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

[96] G. A. Soulodre, N. Popplewell, and J. S. Bradley, "Combined effects of early reflections and background noise on speech intelligibility," *J. Sound Vib.* vol. 135, no. 1, pp. 123–133, 1989.

[97] G. A. Soulodre, N. Popplewell, M. C. Lavoie, and J. S. Bradley, "A suggested modification to the Fairbanks rhyme test," *J. Sound Vib.* vol. 123, no. 3, pp. 578–580, 1988.

[98] J. B. Allen, "Short term spectral analysis, synthesis, and modification by Discrete Fourier Transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, no. 3, pp. 235–238, June 1977.

[99] R. A. Curtis and R. J. Niederjohn, "An investigation of several frequency-domain methods for enhancing the intelligibility of speech in wideband random noise," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 602–605, Apr. 1978.

[100] D. B. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 4, pp. 786–794, Aug. 1981.

[101] P. Scalart and J. Vieira Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 629–632, May 1996.

[102] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, vol. 2, pp. 355–358, Apr. 1993.

[103] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[104] J. Berger, R. R. Coifman, and M. J. Goldberg, "A method of denoising and reconstructing audio signals", in *Proc. Int. Comp. Music Conf.*, pp. 344–347, 1994.

[105] J. Berger, R. R. Coifman, and M. J. Goldberg, "Removing noise from music using local trigonometric bases and wavelet packets," *J. Audio Eng. Soc.*, vol. 42, no. 10, pp. 808–818, Oct. 1994.

[106] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Englewood Cliffs: Prentice-Hall, 1989.

[107] S. Shlien and G. Soulodre, "Measuring the characteristics of expert listeners," *The 101st Convention of the Audio Eng. Soc.*, Nov. 1996. Preprint 4339 (E-2)

[108] T. Kasparis and J. Lane, "Adaptive scratch noise filtering," *IEEE Trans. Consumer Electronics*, vol. 39, no. 4, pp. 917–921, Nov. 1993.

[109] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, Englewood Cliffs: Prentice-Hall, 1983.

[110] J. D. Johnston, "A filter family designed for use in quadrature mirror filter banks", in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 291–294, Apr. 1980.

[111] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs: Prentice-Hall, 1993.

[112] P. P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions and perfect-reconstruction techniques," *IEEE ASSP Magazine*, vol. 4, pp. 4–20, July 1987.

[113] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley: Wellesley-Cambridge Press, 1996.

[114] Y. Meyer, *Wavelets - Algorithms and Applications*, Philadelphia: Society for Industrial and Applied Mathematics, 1993.

[115] O. Rioul and M. Vertelli, "Wavelets and signal processing", *IEEE Signal Proc. Mag.*, pp. 14–38, Oct. 1991.

[116] M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, *Wavelets and Their Applications*, Boston: Jones and Bartlett Publishers, 1992.

[117] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, "Signal processing and compression with wave packets", Dept. Math, Yale Univ., 1990, preprint.

[118] K. Brandenburg, Fraunhofer Institut, *Private Communication.*

[119] D. Sinha and A. H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets", *IEEE Trans. Signal Processing.*, vol. 41, no. 12, pp. 3463–3479, Dec. 1993.

[120] D. Tsoukalas, M. Paraskevas, and J. Mourjpoulos, "Speech enhancement using psychoacoustic criteria", in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, vol. 2, pp. 359–362, Apr. 1993.

[121] T. L. Peterson and S. F. Boll, "Acoustic noise suppression in the context of a perceptual model," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 1086–1088, Apr. 1981.

[122] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Processing.*, vol. 39, no. 9, pp. 1943–1954, Sept. 1991.

[123] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation", *J. Audio Eng. Soc.*, vol. 40, no. 12, pp. 963–978, Dec. 1992.

[124] H. Fletcher, "Auditory patterns", *Rev. Mod. Phys.*, vol. 12, pp. 47–65, Jan. 1940.

[125] B. Moore, *Frequency Selectivity in Hearing*, London: Academic Press, 1986.

[126] B. Moore, *An Introduction to the Psychology of Hearing*, London: Academic Press, 1997.

[127] E. Zwicker and R. Feldtkeller, *Psychoacoustique – L'oreille Récepteur D'information*, Paris: Masson, 1981.

[128] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, Berlin: Springer-Verlag, 1990.

[129] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria", *IEEE J. Selected Areas in Commun.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[130] D. E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Perceptual filters for audio signal enhancement", *J. Audio Eng. Soc.*, vol. 45, no. 1/2, pp. 22–36, Jan/Feb. 1997.

[131] N. Virag, "Speech enhancement based on masking properties of the auditory system," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 796–799, May 1995.

[132] A. A. Azirani, R. Le Bouquin, and G. Faucon, "Optimizing speech enhancement by exploiting masking properties of the human ear", in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 800–803, May 1995.

[133] R. J. Beaton, J. G. Beerends, M. Keyhl, and W. C. Treurniet, " Objective perceptual measurement of audio quality", *AES Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 126–152, Sept. 1996.

[134] B. Scharf, "Critical Bands", in *Foundations in Modern Auditory Theory*, New York: Academic Press, 1970.

[135] 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. Amer.*, vol. 66, no. 6, pp. 1647–1652, Dec. 1979.

[136] R. D. Patterson, "Auditory filter shape", *J. Acoust. Soc. Amer.*, vol. 55, no. 4, pp. 802–809, April 1974.

[137] R. D. Patterson, I. Nimmo-Smith, D. L. Weber, and R. Milroy, "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold", *J. Acoust. Soc. Amer.*, vol. 72, no. 6, pp. 1788–1803, Dec. 1982.

[138] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)", *J. Acoust. Soc. Amer.*, vol. 33, no. 2, p. 248, Feb. 1961.

[139] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory filter bandwidths and excitation patterns", *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 750–753, Sept. 1983.

[140] Model Proponents: ITU-R Task Group 10/4, *Method for Objective Measurement of Perceived Audio Quality-Description of the Models*, Jan. 1998.

[141] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective evaluation of state-of-the-art 2-channel audio codecs", *J. Audio Eng. Soc.*, vol. 46, no. 3, pp. 164–177, March 1998.

[142] M. J. Penner, "The coding of intensity and the interaction of forward and backward masking", *J. Acoust. Soc. Amer.*, vol. 67, no. 2, pp. 608–616, Feb. 1980.

[143] M. J. Penner and R. M. Shiffrin, "Nonlinearities in the coding of intensity within the context of a temporal summation model", *J. Acoust. Soc. Amer.*, vol. 67, no. 2, pp. 617–627, Feb. 1980.

[144] R. A. Lutfi, "Additivity of simultaneous masking", *J. Acoust. Soc. Amer.*, vol. 73, no. 1, pp. 262–267, Jan. 1983.

[145] R. A. Lutfi, "A power-law transformation predicting masking by sounds with complex spectra", *J. Acoust. Soc. Amer.*, vol. 77, no. 61, pp. 2128–2136, June 1985.

[146] D. M. Green, "Additivity of masking", *J. Acoust. Soc. Amer.*, vol. 41, no. 6, pp. 1517–1525, Apr. 1967.

[147] L. E. Humes and W. Jesteadt, "Models of the additivity of masking", *J. Acoust. Soc. Amer.*, vol. 85, no. 3, pp. 1285–1294, March 1989.

[148] B. Paillard, *Codage perceptuel des signaux audio de haute qualité*, Ph.D. Thesis, Université de Sherbrooke, 1992.

[149] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne, "PERCEVAL: perceptual evaluation of the quality of audio signals", *J. Audio Eng. Soc.*, vol. 40, no. 1/2, pp. 21–31, Jan/Feb. 1992.

[150] E. A. G. Shaw, "Transformation of sound pressure level from the free field to the eardrum in the horizontal plane", *J. Acoust. Soc. Amer.*, vol. 56, no. 6, pp. 1848–1861, Dec. 1974.

[151] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals", *J. Acoust. Soc. Amer.*, vol. 71, no. 3, pp. 679–688, March 1982.

[152] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data", *Hearing Res.*, vol. 47, pp. 103–138, 1990.

[153] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", *J. Acoust. Soc. Amer.*, vol. 68, no. 5, pp. 1523–1525, Nov. 1980.

[154] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the basilar membrane", *J. Acoust. Soc. Amer.*, vol. 33, no. 10, pp. 1344–1356, Oct. 1961.

[155] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: a revised scale", *The Amer. J. of Psych.*, vol. 53, no. 3, pp. 329–353, July 1940.

[156] B. C. J. Moore and B. R. Glasberg, "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns", *Hearing Res.*, vol. 28, pp. 209–225, 1987.

[157] J. P. Egan and H. W. Hake, "On the masking pattern of a simple auditory stimulus", *J. Acoust. Soc. Amer.*, vol. 22, no. 5, pp. 622–630, Sept. 1950.

[158] R. H. Ehmer, "Masking patterns of tones", *J. Acoust. Soc. Amer.*, vol. 31, no. 8, pp. 1115–1120, Aug. 1959.

[159] J.J. Zwislocki "Masking: experimental and theoretical", in *Handbook of Perception* vol. IV, edited by E. C. Carterette and M. P. Friedman, New York: Academic Press, 1978.

[160] R. D. Patterson, "Auditory filter shapes derived with noise stimuli", *J. Acoust. Soc. Amer.*, vol. 59, no. 3, pp. 640–654, March 1976.

[161] M. J. Shailer, B. C. J. Moore, B. R. Glasberg, N. Watson, and S. Harris, "Auditory filter shapes at 8 and 10 kHz", *J. Acoust. Soc. Amer.*, vol. 88, no. 1, pp. 141–148, July 1990.

[162] B. C. J. Moore, R. W. Peters, and B. R. Glasberg, "Auditory filter shapes at low centre frequencies", *J. Acoust. Soc. Amer.*, vol. 88, no. 1, pp. 132–140, July 1990.

[163] A. J. Oxenham and B. C. J. Moore, "Modeling the additivity of nonsimultaneous masking", *Hearing Res.*, vol. 80, pp. 105–118, 1994.

[164] W. Jesteadt, S. P. Bacon, and J. R. Lehman, "Forward masking as a function of frequency, masker level, and signal delay", *J. Acoust. Soc. Amer.*, vol. 71, no. 4, pp. 950–962, April 1982.

[165] B. C. J. Moore and B. R. Glasberg, "Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise", *J. Acoust. Soc. Amer.*, vol. 73, no. 4, pp. 1249–1259, April 1983.

[166] B. C. J. Moore, "Additivity of simultaneous masking, revisited", *J. Acoust. Soc. Amer.*, vol. 78, no. 2, pp. 488–494, Aug. 1985.

[167] J. A. Candahl, "Two- versus four-tone masking at 1000 Hz", *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 471–474, Aug. 1971.

[168] D. A. Nelson, "Two-tone masking auditory critical bands", *Audiology*, vol. 18, pp. 279–301, 1979.

[169] R. D. Patterson and I. Nimmo-Smith, "Off-frequency listening and auditory filter asymmetry", *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 229–246, Jan. 1980.

[170] R. C. Bilger, "Additivity of different types of masking", *J. Acoust. Soc. Amer.*, vol. 31, no. 8, pp. 1107–1109, Aug. 1959.

[171] R. H. Wilson and R. Carhart, "Forward and backward masking: interactions and additivity", *J. Acoust. Soc. Amer.*, vol. 49, no. 4, pp. 1254–1263, Apr. 1971.

[172] G. Widin and N. F. Viemeister, "Masker interaction in pure-tone forward masking", *J. Acoust. Soc. Amer.*, vol. 68, no. 2, pp. 475–479, Aug. 1980.

[173] J. D. Johnston, D. Sinha, S. Dorward, and S. R. Quackenbush, "AT&T perceptual audio coding (PAC)", *AES Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 73–82, Sept. 1996.

[174] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Y. Oikawa, "ISO/IEC MPEG-2 advanced audio coding", *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814, Oct. 1997.

[175] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform", *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[176] G. A. Davidson, "Digital audio coding: Dolby AC-3", Ch. 41 in *The Digital Signal Processing Handbook*, V. K. Madisetti and D. B. Williams Eds. Boca Raton: CRC Press, 1998.

[177] L. D. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, and S. Vernon, "AC-2 and AC-3 low-complexity transform-based audio coding", *AES Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 54–72, Sept. 1996.

[178] ISO/IEC 13818-7, "Information technology - generic coding of moving pictures and associated audio, Part 7: advanced audio coding", Dec. 1996.

[179] Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems, ITU-R Recommendation BS.1116, 1994.

[180] T. Grusec, L. Thibault, and R. Beaton, "Sensitive methodologies for the subjective evaluation of high quality audio coding systems", *Proceedings of the AES UK DSP conference*, pp. 62–76, Sept. 1992.

[181] T. Grusec, L. Thibault, and G. Soulodre, "Subjective evaluation of high quality audio coding system: Methods and results in the two-channel case", *The 99th Convention of the Audio Eng. Soc*, Oct. 1995, Preprint 4065.

[182] *Low Bit Rate Audio Coding*, ITU-R Recommendation BS.1115, 1992.

[183] W. A. Gardner, "Exploitation of spectral redundancy in cyclostationary signals," *IEEE Signal Processing Magazine*, pp. 14–36, April 1991.

[184] J. M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications," *Proc. IEEE*, vol. 79, no. 3, pp. 278–305, March 1991.

[185] M. J. Hinich and H. Messer, "On the principal domain of the discrete bispectrum of a stationary signal," *IEEE Trans. Signal Proc*essing, vol. 43, no. 9, Sept. 1995.

[186] J. C. Marron, P. P. Sanchez, and R. C. Sullivan, "Unwrapping algorithm for least-squares phase recovery from the modulo 2π bispectrum phase," *J. Opt. Soc. Am.* vol. 7, no. 1, Jan. 1990.

[187] T. Matsuoka and T. J. Ulrych, "Phase estimation using the bispectrum," *Proc. IEEE*, vol. 72, no. 10, pp. 1403–1411, Oct. 1984.

[188] S. Seetharaman and M. E. Jernigan, "Speech reconstruction based on higher order statistics," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 703–706, Apr. 1988.

[189] I. J. Gdoura, P. Loizou, and A. Spanias, "Speech processing using higher order statistics," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 160–163, Apr. 1993.

[190] R. Fulchiero and A. S. Spanias, "Speech enhancement using the bispectrum," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 488–491, Apr. 1993.

[191] B. Boyanov, S. Hadjitodorov, and T. Ivanov, "Analysis of voice speech by means of bispectrum," *Elect. Letters*, vol. 27, no. 24, Nov. 1991.