The Removal of Environmental Noise in Cellular Communications by Perceptual Techniques.

Mark Tuffy

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
December 10, 1999

Abstract

This thesis describes the application of a perceptually based spectral subtraction algorithm for the enhancement of non-stationary noise corrupted speech. Through examination of speech enhancement techniques, explanations are given for the choice of magnitude spectral subtraction and how the human auditory system can be modelled for frequency domain speech enhancement. It is discovered, that the cochlea provides the mechanical speech enhancement in the auditory system, through the use of masking. Frequency masking is used in spectral subtraction, to improve the algorithm execution time, and to shape the enhancement process making it sound natural to the ear.

A new technique for estimation of background noise is presented, which operates during speech sections as well as pauses. This uses two microphones placed on opposite ends of the cellular handset. Using these, the algorithm determines whether the signal is speech, or noise, by examining the current and next frames presented to each microphone. This allows operation in non-stationary conditions, as the estimation is calculated for each frame, and a speech pause is not required for updating. A voting decision process decides the presence of speech or noise which determines which microphone the estimation is calculated from.

The importance of an accurate noise estimate is highlighted with a new technique to reduce the effect of musical noise artifacts in the processed speech. This is a classic drawback of spectral subtraction techniques, and it is shown, that the trade off between noise reduction and speech distortion can be extended by this process. A new method for dealing with musical noise is described, which uses a combination of energy and variance examination of the spectrogram to segregate potential musical noise from desired speech sections. By examination of the spectrogram points surrounding musical noise sections, perceptually relevant values replace the corruption leading to cleaner enhanced speech.

Any perceptual speech system requires accurate estimates of the clean speech masking thresholds, to prevent noisy sections being passed through the enhancement untouched. In this thesis, a method for the calculation of the estimated clean speech masking thresholds is derived. Classically, this requires an estimation of the clean speech before the thresholds can be derived, but this results in inaccuracy due to the presence of musical noise and spectral nulls. The new algorithm examines the thresholds produced by the corrupted speech, and the background noise, and from these determines the relationship between the two, to produce an estimate of the clean thresholds, with no operation performed on the actual speech signal. A discrepancy is found between the results for male and female speech, which, by examination of the perceptual process, is shown to be due to the different formant positions in male and female speech.

Following the development of these parts, the entire enhancement algorithm is tested on a range of noise scenarios, using male and female speech. The results show, that the proposed algorithm is able to provide adequate performance in terms of noise reduction and speech quality.

Declaration of originality

I hereby declare that the research recorded in this thesis, and the thesis itself, was composed and originated entirely by myself in the Department of Electronics and Electrical Engineering at The University of Edinburgh.

The software routine which produced the spectral subtraction images was provided by Nick Walton.

Mark Tuffy

Acknowledgements

There are many people I would like to thank for helping me produce this thesis, namely Kenny Dalglish, George Lucas, Eric Clapton, Gene Rodenberry and Stephane Adam. Unfortunately, none of them helped in any way with the work here, so I won't. On the other hand, there have been a number of people who have helped, prodded, pushed and commiserated with me over the last few years.

I'd like to thank my supervisor, Dr David Laurenson, who provided many hours of intellectual and practical guidance, and who by now, will be wishing he never proof-read this thesis. His support and advice have been invaluable. Thanks also go to Dr Bernard Mulgrew, who took time out from a very busy schedule to read over this work.

I would like to thank BT Labs, for providing the data which was used in the experiments. I'm sure, that like many others the test phrases used will be for evermore burned into the conscience of those who have heard them.

While he may not realise it, Iain Mann has been a great help in showing me the error of my ways numerous times, and has been a great sounding board for ideas. He also happens to play a mean game of office table tennis and office cricket, which helped while away many a stressed afternoon. Also in the SASG group, I would like to thank Nick Walton and Damon Thomson, who have always been a source of great amusement, Damon in particular shall always stick in my mind for his uncanny Michael Cain and Sid James impersonations. Nick on the other hand, always managed to raise a smile when engaged in "audio evangelism" mode, and his flexible working hours always made me laugh.

A very special thank you goes to those people who have had to deal with me over the years when I left the office and returned home. My wife Rachel has been a tower of strength giving support and encouragement, which has helped me make it through these last three years, also my parents and sister have shown never-ending support and encouragement.

As a final thought I'd like to thank the Hearts Cup Winning team of 1998 for at least allowing me to see Hearts win something in my lifetime.

Thank you all and remember as a famous man once said..."Doh!"

Contents

List of figures

List of tables

Acronyms

ANC	Adaptive Noise Cancellation
LMS	Least Mean Squared
MMS	Minimum Mean Squared
RAP	Row Action Projection
MSC	Magnitude Squared Coherence
CB	Critical Band
BM	Basilar Membrane
PSD	Power Spectral Density
DFT	Discrete Fourier Transform
DCT	Discrete Cosine Transform
WT	Wavelet Transform
STDFT	Short Time Discrete Fourier Transform
SRT	Speech Reception Threshold
SS	Spectral Subtraction
SNR	Signal To Noise Ratio
NEAH	Noise Estimation With a Hangover
PANE	Parallel Noise Estimation
SPL	Sound Pressure Level
LPC	Linear Predictive Coding
MOS	Mean Opinion Score
POI	Point Of Interest
FBLP	Forward and Backward Linear Prediction
CSMT	Clean Speech Masking Threshold

Nomenclature

n	discrete time
$s(n)$	time domain clean speech
$y(n)$	time domain corrupted speech
$n\ (n)$	time domain noise
$\widehat{s}(n)$	estimated time domain clean speech
$h(n)$	discrete time domain hamming window
m	discrete frequency
N	FFT, window size
$S(m)$	frequency domain clean speech spectrum
$Y(m)$	frequency domain corrupted speech spectrum
$\widehat{N}(m)$	frequency domain estimated noise spectrum
$\widehat{S}(m)$	frequency domain estimated speech spectrum
$\|S(m)\|$	magnitude spectrum
$\theta(m)$	phase spectrum
$\|S(m)\|^2 = P(m)$	power spectrum
$\epsilon(m)$	spectral error
α	power exponent
β	oversubtraction factor
η	noise floor percentile
k	critical band number
j	center point of critical band
$B_k(j)$	critical band filter
a_k	lower limit of critical band
b_k	upper limit of critical band
$Sp(k)$	perceptual spreading function
B_k	bark spectrum
C_k	critical bark spectrum
T_k	estimated masking threshold

E_{prim}	primary microphone energy
E_{ref}	reference microphone energy
p	forgetting factor
$\mu(m)$	average value of the background noise
r	x axis point for spectrogram
t	y axis point for spectrogram
i	frame number
τ	energy multiplication factor
$E_A(r,t)$	energy of spectrogram region A
$E_B(r,t)$	energy of spectrogram region B
B_c	blade
c	blade number
B_{min}	minimum variance blade
$a1, a2$	y axis boundary points of region
$b1, b2$	x axis boundary points of region
b_{sp}	spectrogram points
sp	spectrogram point index
w_M	forward linear prediction tap weight
M	number of tap weights
$u(n-M)$	tap inputs
$\widehat{u}(n)$	linear prediction output
g_M	backward linear prediction tap weight
$\mathbf{R}$	correlation matrix of the tap inputs
$\mathbf{r}$	the cross correlation matrix
δ	threshold energy subtraction factor
MAP	minimal audible pressure
$diff_{th}$	threshold difference
$Th_{Y(m)}$	corrupted speech threshold
$Th_{\widehat{N}(m)}$	estimated noise threshold

Chapter 1
Introduction.

1.1 Implementing speech enhancement into mobile phones

One of the largest growing consumer markets in the 1990s is in the mobile communications industry. Technology in this area has progressed so rapidly, that the cumbersome heavy handsets with limited battery life first introduced in the late 1980s have been replaced by small lightweight long life devices, which can also incorporate fax and email facilities. So far, the only limit to the size of such handsets, are the physical dimensions of the keypad, which if shrunk too far, would become impractical to use.

The drive towards a cellular system which gives worldwide coverage is the next goal. This opens up the possibility of a customer being able to use the same mobile phone almost anywhere on the planet. Given such freedom, the number of locations a handset can be used in is vast, and each location will have its own particular noise environment. Whether it be walking along a busy city street, waiting in a train station, sitting on a bus, or trading on a stock market floor, the environmental noise will continually change. In such cases, the quality of speech the listener receives can be poor, and difficult to understand. If the background noise can be reduced so that it does not interfere with the speech, the ability to hold a conversation would be greatly enhanced. This is the reason why speech enhancement is a topic of great interest for the cellular operators.

Research has been conducted into mobile phone speech enhancement, but it has tended to concentrate on the area of hands free systems for cars [4] [5], which rely on the fact that the acoustic environment of a car can be easily modelled. Little work has been presented which looks at the scenario of a user moving through an environment, where there are a variety of varying noise conditions.

One branch of research falls under the category of spectral subtraction, which on its own, can produce limited results by operating on the magnitude spectra of the signal. With the addition of extra elements, it can provide acceptable speech quality for the mobile phone problem. Of

particular interest in this respect, is the ability of the human auditory system to filter out aspects of background noise that are of no interest, allowing the user to concentrate on that of use. This is a technique which goes on subconsciously, with only extreme cases requiring the listener to concentrate on a signal of interest. The integration of such an element into speech enhancement would be of great benefit, especially in very low signal to noise ratio (SNR) environments.

1.1.1 Aims of the thesis

The aim of the work presented here is to design a speech enhancement system which, can remove non-stationary environmental noise from a mobile phone conversation. Such a system must produce speech that is easy to understand, and of a good quality, as perceived by the listener. This is of particular importance, as the user will perceive a good quality service as one that provides a clear easy to understand conversation. This system is a preprocessor technique, and would be implemented before the data is coded by the handset codec.

The algorithms presented here introduce new methods for;

- the production of a noise estimate for spectral subtraction, which can operate both during speech and pause sections, using dual microphones to produce as close an estimate as possible. This operates on a very short time scale (16ms windows of the signal), in order to ensure accurate estimation of the noise during non-stationary conditions.
- the removal of unwanted artefacts commonly known as “musical noise”, which sounds like a warbling or watery effect on the enhanced speech. This allows the tradeoff between noise reduction and speech distortion to be stretched, and maintain the quality of the processed speech.
- the calculation of clean speech masking thresholds for use in the perceptual spectral subtraction, without exposing the threshold calculation to speech distortions.

An important point for the mobile case is ensuring that an accurate noise estimation is provided. For the non-stationary condition which is being assumed, this requires not only a software approach but some thought into the sensors, which would provide the inputs. The positioning of the reference microphone to detect the majority of the background noise, and the primary microphone to detect the corrupted speech, and any changes between the noise presented to the speech and the reference, are important. This is a physical problem, which must also be

addressed, in order to justify the software approach taken, and as such, is not addressed in this thesis.

1.2 Thesis layout

The purpose of this work, is to develop a perceptually based speech enhancement system for mobile phone use. Subsequent chapters show the process behind the development of the algorithm.

Chapter 2 describes some of the typical speech enhancement algorithms used in the past and present. The topic of adaptive filtering, its utilisation in the areas of active noise, and echo control are presented. Hearing aid wearers suffer from the same problem of hearing speech in noise as the mobile phone case, and the technique of beamforming is examined as a possible solution for this.

The use of spectrally based algorithms is a popular field in enhancement. As the presented work falls under this category, a background into various spectral techniques such as RASTA processing introduces the concept. Work done by Boll [6] introduces the concept of spectral subtraction, and the various variants which can be used. The perceptual aspect of the algorithm is shown by examining the functions of the human auditory system. By breaking the system down into constituent parts, it is possible to determine the role of each, showing which is useful for speech enhancement. Lastly in Chapter 2, it can be seen that for spectral subtraction, there is the need for a signal to remove artefacts from the corrupted speech. The importance of an accurate noise estimate is shown, which will help deal with any tonal artefacts that may occur in the speech after enhancement.

Having chosen to use the spectral subtraction method, Chapter 3 expands on this, to explain the variant chosen. The choice is determined by a number of parameters, such as the percentage of reference signal magnitude to be subtracted, whether power or magnitude spectral subtraction is appropriate, and how to deal with negative values in the spectrum. Negative spectral values can occur in the processed speech, as the interaction between the speech and noise in the corrupted signal can be either constructive or destructive. If the interaction is destructive, then the corrupted speech magnitude may have a lower sound pressure level than the estimated background noise. When the noise is subtracted from the speech magnitude, this would result in a negative value. These points cannot be left in the enhanced speech, as they will lead to

"musical noise". For the transformation into the spectral domain, the choice of data window and overlapping are also important in speech. Once these have been made, the framework for the enhancement is in place. The perceptual aspect comes from introducing masking thresholds as produced by the cochlea. Using these, it is possible to increase the speed of the algorithm, by only choosing audible points, which are suitable for enhancement.

Having established in Chapter 2 the importance of an accurate noise estimate, the next step is ensuring that the spectral subtraction algorithm manages to obtain as good an estimate as possible. The signals from the primary and reference microphones are examined frame by frame, to determine if the primary contains speech or noise. As the algorithm is to work in a non-stationary environment this is an important step. The *Noise Estimation with A Hangover* (NEAH) algorithm, reveals how waiting will confirm whether a frame is actually noise, and this can help the estimate. An interesting point, is that there are only a small number of noise estimation algorithms. Each one may perform well in a particular situation, and poorly in others. Rather than take only one, it is shown that running these in parallel, would produce a better performance. This leads to the *Parallel Noise Estimation* (PANE) algorithm, whose performance is compared to that of the algorithms which are included in it.

With the estimation of the background noise, and the speech having passed through the auditory spectral subtraction, the first step of the enhancement process has been completed. There is now the need for a system to remove any unwanted artefacts, which the enhancement process may have caused. Chapter 5 details the effects of this so called "musical" noise, and the subsequent reduction in speech quality. By examining the spectrograph of the speech, it is possible to view the difference between the original clean speech, and the enhanced speech. To counter this the spectrogram is scanned in both time and frequency, to determine which points are out of place. Through a mixture of energy tracking, variance comparisons, and linear prediction, anomalous points are removed, and speech points which have been erroneously deleted, are restored. Listening tests show the performance of the algorithm in a perceptual manner, and these are presented along with statistical and graphical results.

In the perceptually based spectral subtraction routine, the masking thresholds are critical in determining which points are suitable for enhancement, and which are not. If a good estimate of the clean speech masking thresholds are not used, the corruption to the resultant speech can be significant. Chapter 6 introduces a new approach to the estimation of the masking thresholds, without performing any operation on the speech itself. This allows the production of thresholds,

without the influence of any distortion from a subtraction process.

The conclusion of the thesis is Chapter 8 which brings together the results of all the work presented in the previous Chapters, along with the final listening tests presented in Chapter 7. The merit of the algorithm produced are discussed in terms of the feasibility of its implementation, and the quality of the results produced. A discussion on the listening tests in Chapter 7 allows an impartial gauge of the speech quality produced by the enhancement process, and points to areas where future work could improve performance.

1.3 Presentation of results and examination of spectrograms

Unless otherwise stipulated, all of the results presented in this thesis in either graphical or tabular format, are the result of experimentation carried out by the author. All of the software was written in C, compiled, and run on a Sun Ultra 10 workstation. The test data operated on, was 16 bit, telephone bandwidth ($0 - 4$kHz), and sampled at 8kHz.

In order to graphically display the effects of the differing algorithms applied to enhance the speech, a number of spectrograms are presented. An example of this is shown in Figure 1.1.

In Figure 1.1(c), it can be seen that the speech after application of an enhancement process differs from the ideal clean speech given in Figure 1.1(a). Sections which cause musical noise are highlighted, along with speech sections which have been erroneously removed, causing speech distortion. It can also be seen, that there is a reduction of the corrupting noise which was applied to the signal in Figure 1.1(b). The purpose of the spectrograms, is to show the comparative performance of algorithms as the amount of noise reduction, musical noise, and speech distortion which occurs can be seen, and compared.

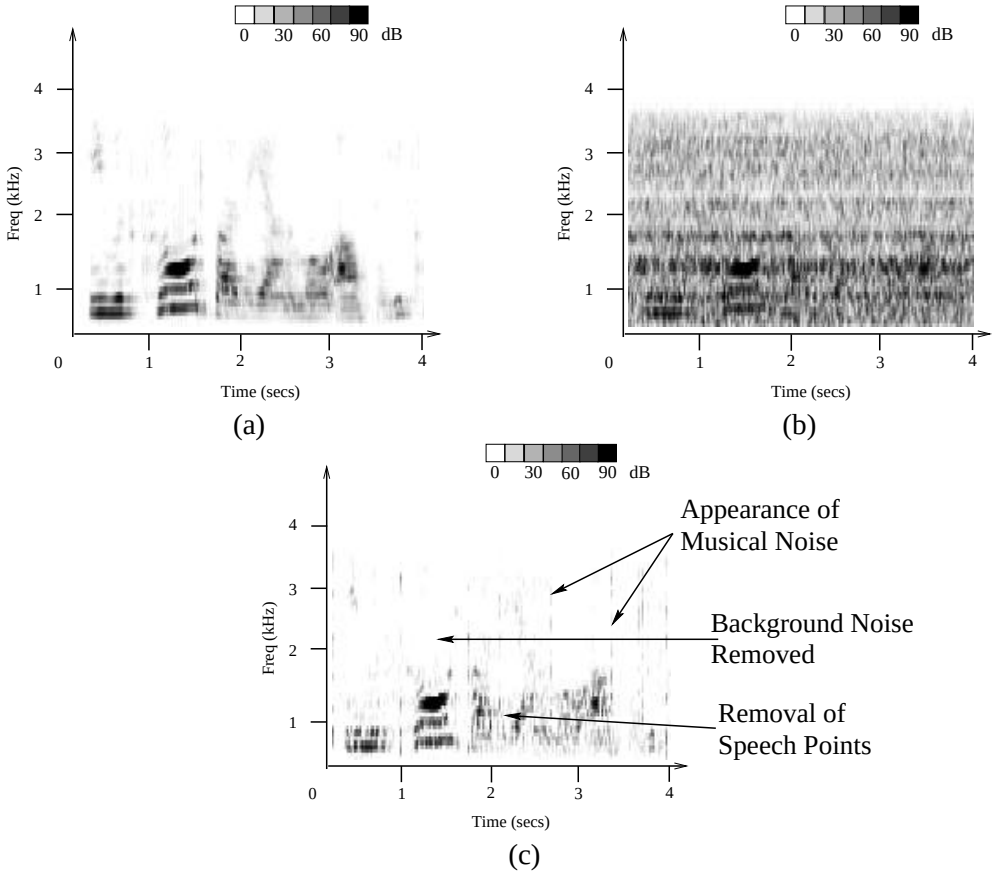


Figure 1.1: *Spectrograms of (a) Clean Female Speech (b) Corrupted speech (c) Speech after enhancement*

Chapter 2
Speech Enhancement Background.

In this Chapter, an examination of the background of some techniques used in speech enhancement is given. The concepts of adaptive filtering and beamforming are introduced, before moving on to spectral techniques. Some of the techniques used in the frequency domain to enhance speech are shown, before moving into the background of spectral subtraction, and discussion of perceptual techniques. The basis behind the calculation of an accurate noise estimate, and the effects of musical noise, are introduced. Each section will contain a brief overview of the work which has been presented in these fields.

2.1 Techniques for speech enhancement

In Chapter 1 the task of enhancing mobile phone speech in a non-stationary environment was introduced. In order to develop an approach which would be applicable to this scenario, it was necessary to look at the problem of speech enhancement to see how this had been tackled.

The first assumption made, was that the only signals available to operate on were the corrupted speech, and an estimate of the background noise. From this, it became obvious that the situation was ideal for the application of adaptive filtering, which could possibly suppress or cancel out the background noise.

2.2 Adaptive filtering

The topic of adaptive filtering has long been used in an attempt to enhance signals of various kinds, not only speech. In [7], some of the earliest known work on adaptive noise cancellation (ANC) was presented. In general the application of ANC algorithms requires the use of two sensors, a primary which receives the corrupted speech input, and a secondary for the reference noise source. In a non stationary environment, single sensor systems can be used for adaptive filtering, but they can only perform echo cancellation effectively.

2.2.1 LMS based ANC

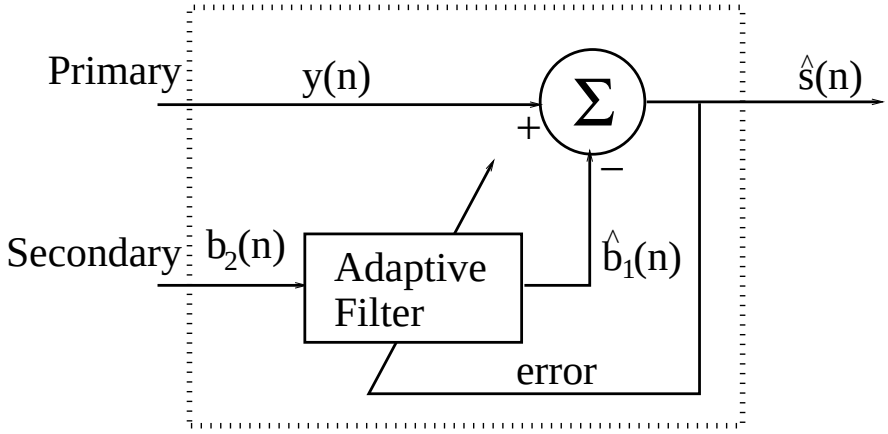


Figure 2.1: *General ANC.*

Figure 2.1 shows the general layout for adaptive filter based ANC. The primary sensor receives an input signal $y(n)$, which is a combination of the desired signal, and corrupting noise. The secondary sensor receives a noise input $b_2(n)$, which is uncorrelated to the desired signal. This is passed through the adaptive filter, in an attempt to estimate as close as possible the noise added to the primary channel $\widehat{b}_1(n)$. This estimate is subtracted from the primary input $y(n)$, and the error fed back into the adaptive filter, to update the filter coefficients.

The main criterion for the adaptive filtering is choosing the algorithm to determine the manner in which the filter coefficients are updated. The most common of these is the Minimum or Least Mean Squares Error (MMSE, LMSE) estimation, which minimises the mean power of the output. The adaptive filter is implemented as a finite impulse response (FIR) design, with the most common Least Mean Squares (LMS) algorithm the Widrow-Hopf LMS algorithm normally used, as shown in [8]. LMS based algorithms were used in [9] and [10] as methods to remove noise from aircraft cockpit speech.

In both [10] and [11] it was proposed that the primary microphone was inside the pilots oxygen mask. The reference microphone was placed on the outside of the mask, enabling it to record the ambient noise of the cockpit. One of the main drawbacks of ANC, is that the sensors must be well separated, to prevent any speech cross talk in the secondary input, which would result in the removal of some speech signal. This is important, as there is a proportional increase between the separation of the sensors, and the number of filter taps that are required. This is

due to the delay that occurs between the reference, and the primary microphone inputs. These longer filter lengths lead to an increase in the time required to calculate the coefficients and filter the signal.

One of the main disadvantages to the implementation of LMS based adaptive filtering for speech enhancement, is the time it takes for the filter tap weights to converge to the optimal values. It was stated in section 1.1.1 that the enhancement algorithm would be operating on speech segments in the order of 16ms, and it is not possible for LMS based adaptive filtering to converge in that time. If the noise field is non-stationary, then the filter coefficients would not be able to converge quickly enough to the optimal set for the present noise conditions, during which time the noise conditions have changed, meaning the coefficient set could be non-optimal.

LMS based ANC was used in [12], to reduce both noise and echo in phone conversations, with both single and dual sensor systems presented. In the single microphone case, the echo cancellation was achieved by using LMS with linear prediction, and an adaptive step size controller to reduce the convergence time. A second LMS adaptive filter prevented distortion of the speech during echo removal in double talk conditions.

It was stated in [12], that the optimal spacing of the microphones for noise reduction is approximately $0.4 - 0.5$m. At this point, noise components within each microphone signal could be assumed to be uncorrelated. Results of 10 to 15dB improvement were reported for adaptive filters with 3000 taps for the single microphone case, with noise reduction and echo cancellation of approximately 15dB for the dual case, with no audible residual echos.

In [13], an overview was presented of the development of noise reduction and echo cancellation systems. It was suggested, that the problem of convergence for LMS based adaptive filters can be overcome by pre-whitening the inputs, or by adaptive step size control. If the convergence time was still too long, then a number of different algorithms were presented which could be substituteed for the LMS, such as the Fast Affine Projection, or the Recursive Least Squares. It was stated though, that while such alternative algorithms may lead to reduced convergence times, they did have drawbacks of their own, mostly the instability of the filter tap weights.

The classical systems for noise and echo control relied on a cascade of two separate optimal filters. Recent work has centred on attempting to reduce this complexity, by producing a single filter which could do both jobs, the work presented in [14] and [15] are examples of this.

In [14], a global Wiener filter in the frequency domain was implemented, which was claimed to reduce the computational complexity compared to the cascade algorithm. This was tested by applying signals containing the same speech, but differing car cabin environments. It was claimed that the computational load of this global filter was 80 times lower than that of the traditional cascade design, while producing equally good results.

It can be seen that combined noise/echo reduction systems are possible, through the use of adaptive filtering. The relatively narrow range of applications to which they have been applied was noted. All of the applications studied here for adaptive filtering are applied in acoustic environments, which can be easily modelled and are stationary in nature. Stationary conditions allow the adaptive filter systems the time to converge to an optimal set of filter taps. Once again, if the noise conditions are non-stationary, as in the proposed speech enhancement algorithm, then the adaptive filters would be unable to converge to an optimal set of coefficients in the 16ms time window, meaning that the noise reduction would not operate optimally. This was a condition examined in [16], where the problem of slow convergence was tackled.

2.3 Multiple sensor and Beamforming

The investigation into the techniques of ANC have pointed to one particular aspect of noise reduction, the number of sensors used. It can be seen that in general, applications of ANC require more than one sensor in order to achieve noise reduction as well as echo cancellation, with nominally two or more being used. With a multiple number of sensors, the system becomes a microphone array, with the possibility of beamforming to localise signals in much the same way as radar pinpoints targets. Some excellent background into beamforming and its applications could be found in [17], which gave descriptions of the many types of beamforming such as the multiple sidelobe canceller. A typical beamformer layout can be seen in figure 2.2

In figure 2.2 the microphone inputs $y_1(n) \rightarrow y_L(n)$ are multiplied by associated tap weights $tw_1 \rightarrow tw_L$ to shape the beampattern, and the filters compensate for the propagation delays between the microphone inputs. At the simplest level of two sensors, beamforming corresponds to the setup of the human ears, and this led to the idea of looking at how people who use hearing aids cope in noisy situations.

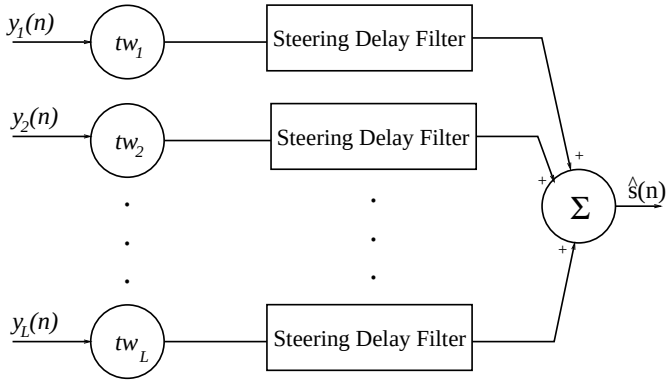


Figure 2.2: *Classical Broadband Beamformer*

2.3.1 Hearing Aid noise reduction

In [18], it was said that approximately 7.5% of the US population have some form of hearing loss, which requires the use of a hearing aid. One of the major problems for such people is the ability to pick out a specific voice in a noisy area, the so called *"Cocktail Party Effect"*. This is due to the lack of binaural detection in the microphone used by hearing aids, causing noise signals to exhibit a higher degree of interference to the hearing aid user. This pointed to a remark in [18], where it was claimed that the SNR increase required to improve the intelligibility of speech in noise for a hearing aid wearer, was far greater than for a normal hearing person.

A two microphone beamforming technique for hearing aid wearers was introduced in [1]. The premise was to try and provide a directional field of listening, which enhanced sounds that were face on, but suppressed sound from other directions. The technique involved the placing of a microphone by each ear of the listener as shown in figure 2.3, which received a mixture of desired and unwanted sounds. By transformation into the frequency domain by the FFT, a noise reducing gain was produced by examining the phase signals which appeared at each microphone. When multiplied with the left and right microphone signals, this produced an approximation to the desired speech. The gain was produced by a recursive system similar to work in [19], by examining the magnitude and phase angles of the signals applied to the microphones. The gain function provided no attenuation if the subject lay perfectly in the centre of the left and right ears, causing the phase angles to be identical in each microphone. Otherwise the difference of the phase angles caused the gain function to attenuate the signal.

In [1], the authors claimed that the results from subjective listening tests performed by hearing

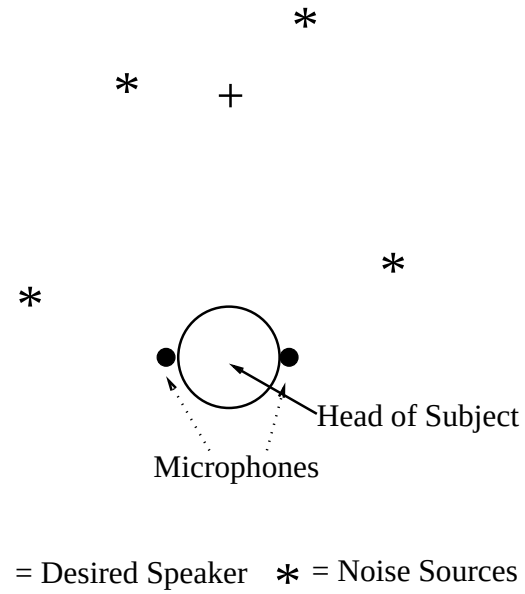


Figure 2.3: *The basic microphone arrangement of [1].*

aid wearers showed that their beamforming algorithm produced speech that was far easier to understand, with approximately a 7dB improvement in SNR. The tests were only performed over the range $[-3 \rightarrow 3]$dB as below -3 dB, they claimed that there is generally no intelligibility for hearing aid wearers, where as above 3dB it is 100%. The problem which arose from this technique, is that the results claimed at the time could only be achieved through the listener wearing a headset, which placed a microphone just above each ear, a cumbersome arrangement which would not be practical for everyday use for a hearing aid wearer. If the microphones are not positioned in such a way the intra-phase relationships of the microphones would not be applicable to determine which signals are off-center, and must be suppressed, as opposed to on-center and desirable.

In order to try to adapt the outputs of a microphone array into a system more useful for hearing aid users, [16] introduced the idea of combining the inputs of a microphone array down to one signal, which would then be presented to the hearing aid. This utilised an adaptive spatial filter, to attenuate off-centre sounds, and leave on-centre sources unattenuated. The filter attenuation depended on the point at which 50% of the presented speech is unintelligible, (the so called speech reception threshold SRT).

The time domain spatial filter of [16] consisted of a number of microphones, each with multiple tap weights per microphone. Each microphone input passed through an FIR filter, and the coefficients of these were combined to give a snapshot vector of the filter response at the time. A

matrix was produced by stacking a number of vectors, each delayed by a time step, and the array output given by multiplying the input signal with this filter snapshot matrix. A technique called minimum variance beamforming was used to solve for the weights, and a modified version of the LMS algorithm was used, as an adaptive solution for the matrix. The performance of this technique was compared to results obtained from a single microphone placed at the front centre of the head, and pointing at the desired speaker.

One problem with adaptive microphone systems concerns the time it takes to obtain a stable adaption of the filter. Short speech segments can cause transients in the weight vectors, as the input signal changes rapidly. As [16] claimed it took one second for the filter to adapt, an impulsive input would result in the appearance of audible artefacts in the processed speech. In [16] it was attempted to remedy this by altering the gain depending on the short time power of the input speech. Large impulses caused a reduction gain and a drop in the requirement of the adaption process, hopefully leading to a stable solution more quickly.

The type of speech enhancement which can be applied to a hearing aid also depends greatly on the type and severity of the hearing loss that has been suffered by the individual. The application of beamforming does not deal with speech enhancement for a specific form of hearing loss, it is a more general technique to help whatever hearing aid they may already use. There have been schemes [20], [21], which have looked at more specific forms of hearing loss, and tried to apply specialised signal processing techniques, to relieve these symptoms. Unfortunately these are very specialised in their application, and no matter how novel their approach or theory may be, they cannot be applied to the more general speech enhancement scenario.

A situation very similar to that of hearing aid wearer is presented in [22], where a microphone array was used to enhance speech, so that a remote listener with access to only one audible signal, (a telephone ear piece, a loudspeaker or a hearing aid perhaps), could hear clean speech from a situation where there was a great deal of interfering noise sources. A four microphone array was used which could be either concave or rectangular with the desired speaker sitting in the middle of the array The first section of the concave case is shown in figure 2.4.

From figure 2.4, it was assumed that the speech signal will arrive at all the array microphones at exactly the same time. The noise source on the other hand, was assumed to be off centre and this led to the noise being presented to the array microphones at slightly different times. It was assumed that the arrival times of the noise at the microphones was such, that they did

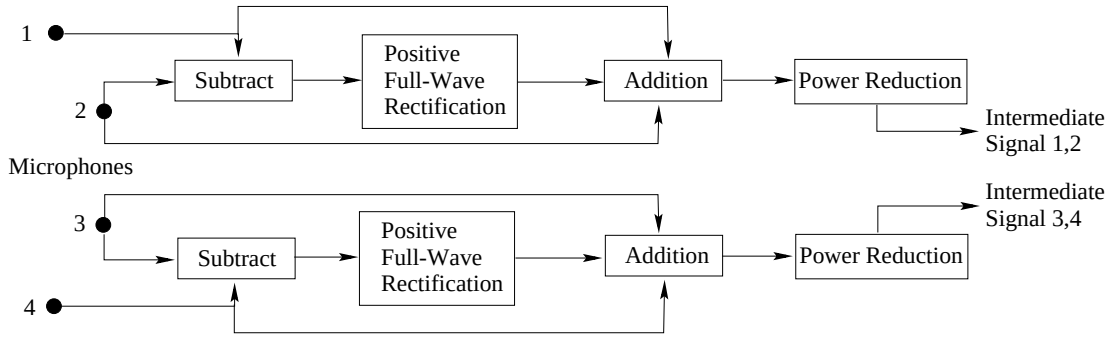


Figure 2.4: *The first section of the process using a concave microphone array*

not overlap temporally. It could be seen in figure 2.4 that the array was split into microphones pairs (1 and $2, 3$ and 4), and these go through subtraction to produce an estimate of the noise. Positive half wave rectification was applied, and the result was added to the original corrupted signals, to give an intermediate signal which was attenuated by 6 dB, to normalise the signal close to the original level.

This first stage represented in figure 2.4, results in two signals (one for each microphone pair), which have a mixture of the speech and the positive peaks of the noise signal not overlapping in time. The operation performed in the second stage of the algorithm was almost identical to the previous, but in this case, only two inputs were presented (the intermediate signals shown in figure 2.4). Also instead of positive half wave rectification, negative rectification was used instead. After this, the signal was added to the intermediates presented to the second stage, and attenuated by a further 6dB. The two stages of rectification which were applied, were an attempt to cancel out both the positive and negative half cycles of the noise signal. As the first stage process only allowed noise with positive peaks to pass, negative half wave rectification should lead to the removal of these peaks, cancelling out the off centre noise corruption.

By examination of the process when signals which do overlap in time were applied, it was discovered that the algorithm in [22] produced an input to the second stage, which was equivalent to the maximum input value of the microphone pairs. For the four microphone inputs, the branch with the largest signal value passes untouched to the input of the second stage. It was found that of these two largest signals, a minimum operation was performed by the second half of the algorithm, resulting in nominally the second or third largest microphone output signal being reproduced at the algorithm output. This meant that conceptually the operation which was produced was equivalent to a maximum-minimum process. A number of tests were performed

comparing the maximum-minimum process to a linear supposition algorithm, and by reversing the order of the rectification operations, a minimum-maximum process. The microphones were arranged so that they enclosed a rectangular space, throughout which the interfering source could be moved. The desired speaker position was always located at the centre of the rectangle. Suppression results of the range $2.6 \rightarrow 5.7$dB for the maximum-minimum process were claimed, as long as the desired speaker position stayed in the centre, although as a rectangular array was used the speaker source can be omnidirectional.

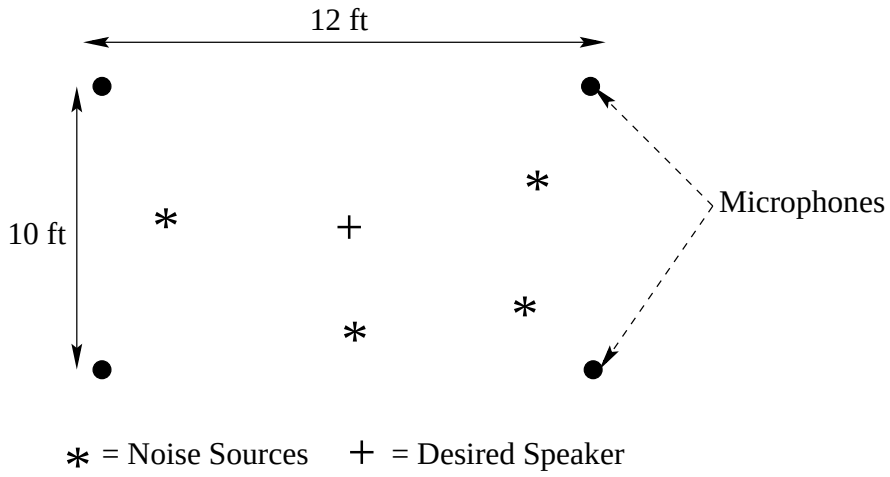


Figure 2.5: *The rectangular four microphone array*

Unfortunately this technique seems to have two drawbacks. Firstly, there is the move from a concave microphone arrangement to a rectangular shape as shown in figure 2.5. The dimensions shown in figure 2.5, are those used in the tests performed in [22]. It can be seen that the microphones sit at the corners of a rectangle, meaning that the technique can only be applied for enhancement of a single speaker in a room, and is not applicable to a hearing aid wearer or a mobile phone user. The dimension of the problem seem to lead it more towards application in conference rooms or speaker-phones, where the user stays fairly stationary. This leads to the second point, which is that for the algorithm of [22], the speaker must be exactly in the centre of the rectangle (as shown in figure 2.5), to ensure that his speech reaches all of the microphones at the same time. Any movement of the desired speaker throughout the room would mirror the movement of a noise source, and lead to the removal of the desired speech from the signal. The performance of this algorithm degrades if this situation occurs, and the simple linear supposition system obtains better results. While it had been suggested in [22] that this could be overcome, by making the desired speaker carry an ultrasonic beacon, allowing the array to calculate the correct gain and delay elements for the speech to be identical in all the

microphone channels, this is not a realistic proposal for many situations.

It can be seen, that the application of beamforming to improve hearing aid performance can produce reasonable results, and there are a number of techniques which can be applied in an attempt to provide a solution to this problem. The main problem for the implementation of these, is the need for more than one sensor or a microphone array. The suggestion that a person wears headgear to give better hearing aid performance is impractical, and some of the solutions presented in this section use microphone arrays, which are of no use to a system designed for a single person. At this point it seems that in order to reduce non stationary noise, a system with at least two sensors may be needed, so it seems prudent to look at other speech enhancement applications of beamforming, to see if they can provide any more suggestions as to a possible methodology.

2.3.2 Other beamforming applications

A great body of work has been presented in the examination of multisensor noise reduction systems for hands-free cellular operation, which centers mostly on the reduction of noise in car environments. Despite the fact that the noise environment in a car is unlike the non-stationary conditions for the proposed mobile phone environment, the work produced on multisensor techniques to solve this problem may provide some information on more practical microphone distances, and enhancement techniques.

The work in [5] is interesting from the point of view that the application is the enhancement of synthesised speech, which may be used in car navigation or traffic announcement systems. The purpose of this work is to render the synthesised speech more intelligible to the driver, allowing important information to be easily heard. The enhancement is based around the Generalised Sidelobe Canceller of [17]. The mainlobe was directed at the synthesised speech, while the adaptive beamformer portion was used to remove the cabin noise. The adaptive beamformer also prevented the synthesised speech from being cancelled out by the adaptive filtering.

The test data was recorded by playing a clean recording of the synthesised speech through a loudspeaker in a car driving at approximately 60mph. The array microphones were placed $40-50$cm away from the loudspeaker. The results obtained came from a range of six different speakers, and improvements of $10-15$dB were recorded.

It was recorded in [5], that there was a problem with the frequency response of the micro-

phone array. As the noise conditions changed, it was noted that the lower frequency range of the speech was lost, and the change could be as high as 20dB. As this algorithm was to be used for speech recognition, this distortion of the spectrum would severely affect the power measurements used in many speech recognition systems. If such a technique was used to enhance non-stationary noise, the possible spectral distortion would result in processed speech that sounded very unnatural.

Moving away from the car cabin environment, [23] presented two microphone array algorithms which aimed to enhance speech corrupted by a second speaker. This was especially of interest, as one of the conditions which a mobile phone user is certain to experience, is the interference of nearby talkers, whether from a PA system in a train system, or other cellular users themselves. [23] enhanced the desired speech, and removed the interfering speech by applying a constrained LMS algorithm and the Row Action Projection algorithm (RAP). It was assumed that the direction, arrival time, and the frequency band of the desired speech were known *a priori*.

The array used consisted of nine microphones, spaced 4cm apart and followed by a 13 point tapped delay line. Two different algorithms were used to solve the adaptive beamforming portion of the algorithm, the first being the Frost algorithm which is a constrained LMS approach solving the LMS optimisation equations via Lagrange Multipliers.

The second method presented was the RAP, method which is an iterative technique for solving linear equations. This expands the LMS optimisation equation:

$$\text{minimisation of } W^T \mathbf{R_{xx}} W \tag{2.1}$$

into a matrix which is a set of linear equations with the constraint of:

$$C^T W = H \tag{2.2}$$

where W, $\mathbf{R_{xx}}$, H, C are the vector of tap weights, the covariance matrix of the received signal, the FIR filter coefficients of the desired frequency response, and the the constraint matrix respectively.

The algorithms were tested on the desired signal normal to the microphone array, and an interfering signal arriving at 30 degrees above the normal. From these test conditions, the average

SNR improvement for the two algorithms was 9.3dB. The angle of the interfering speaker was then swept through from $-90 \rightarrow +90$ degrees and the interference suppression was calculated. From this, it was found that the maximum suppression for the Frost method was 11.2dB and 11.0dB for the RAP method.

[24] implemented the LMS algorithm in the time domain, instead of the frequency domain. This involved the processing of the signal in two distinct steps. The first used the directivity of the beamformer to reduce the noise component in the summed signal by 6dB. This only used three of the four inputs, the fourth being passed through a Wiener filter, to produce the error to train the filter in the second stage. This could be done, as the spacing between the array microphones was such, that the noise was assumed uncorrelated between input channels, meaning that this first filter could only approximate the signal component common to the inputs from the beamformer, which was the speech.

The second stage used a Wiener filter to post-filter the summed signal, which further reduced the noise component. The adaption of the Wiener filter was controlled by the average of the outputs of the microphone array which, as it was less noisy that the actual array inputs, could be used to approximate the clean speech. Tests were performed, using a square microphone array with the desired speaker sitting 0.6m away, and the noise sources being 3 metres away from the array. Unlike some of the adaptive algorithms previously examined, this only required 33 coefficients for the filter, which should result in a reduced convergence time although not as short as the 16ms windowing described in section 1.1.1. The results claimed a $17 - 20$dB improvement in noise level, over an individual array element.

This technique has the advantage of working in the time domain, meaning a reduction in execution time caused by not transforming into the frequency domain. The drawback of this is that, very little extra processing is performed to reduce the noise in the signal leading to an audible residual noise on the speech, which is not critical if the SNR of the desired speech is high. In the conditions which are to be examined for the mobile phone, there is a strong chance that the SNR of the signal will be very low, meaning that such a time domain approach may not be applicable.

2.3.3 Two Microphone systems

In the previous section, it was shown that some speech enhancement solutions use multiple microphone arrays. While under certain circumstances these may give good performance, the general problem is the need to use an array. The size of the arrays required was seen to be at least 0.4m, with some being arranged in a square, some in an arc, and some in a planar field. While this arrangement may work well in cars and rooms, it is of no use for a cellular phone user. In this case it, was decided to look at systems which use a smaller number of sensors.

A technique to reduce the noise found in hands free communications was shown in [25], where it proposed that the system used two microphones, and that each contained a section of speech and noise. It was assumed, that the speech segments of the signal were highly correlated, where as the distance between the microphones was such that the noise was uncorrelated.

The enhancement was obtained by computing the magnitude squared coherence (MSC) from the two channels, and applying this to each channel raised to a different power. This was a technique first introduced in [26]. The purpose of this, was to obtain a weakly distorted estimate in one channel, and a signal reference in the second, which could be used to determine where speech was present in the first channel. For the sections with strong coherence, they would be detected easily by the reference channel, and the lack of coherence in the noise would mean it was not detected. This detection filter was then applied to the first channel, to obtain noise reduced speech.

Test data was collected in a car using microphones which were placed 73 cm apart. This distance was required to ensure that the noise present in each microphone was uncorrelated, allowing the coherence approach to be applied. Informal listening tests indicated that the use of the second channel speech identifier improved noise reduction, compared to the case where the output of the first channel after MSC filtering gave the output.

Another dual sensor technique was presented in [4], which was introduced in Chapter 1 as another in-car noise reduction technique. This used a technique of specifically allocating one microphone to detect the speech and the other to record the environmental noise. This was interesting, as in the previous techniques either ANC or beamforming, have always assumed that all the microphones which were used had both speech and noise in them. In divorcing the recording of the noise from the recording of the speech signal, this allows for the possibility of a more accurate reduction of the noise.

One of the problems of ANC as stated in section 2.2, is the time it can take for convergence of the adaptive filters. To try and get round this, [4] looked at splitting the signal into a section of sub-bands, and cancelling the noise in each subband. This allowed the ability for each sub-band to be operated on in parallel, and as each section was smaller, the time for adaption was greatly reduced. In [4], eight sub-bands were used with filters with $50 - 100$ taps. While the problem of stationarity and non-stationarity with ANC were still relevant, the interesting arrangement of the microphones in this method suggested the possible use of a reference microphone in the mobile phone case.

A dual sensor technique for the separation of simultaneous voices was presented in [27]. The relevance of this work, was dealing with a situation where the competing speakers were of comparable volume. This method exploited the phase difference between the arrival of the signals at the two microphones much in the same way that the human auditory system does to localise sounds.

Once the frequency spectrum of the two microphone signals was calculated, the phase information was derived for each microphone, and from the difference in phase, a number of path differences of a source emitting that frequency were calculated. This information was plotted on a diagram, which had the path difference on the x axis and frequency on the y axis called the *f-p plane*. A point source would show as a vertical line, which corresponded to all the frequencies of the speech coming from the same position. For two speakers who were talking concurrently, although not from the same location, this would result in vertical lines appearing in different parts of the plot.

Neural adaptive windowing [28] was applied to scan over the f-p plane, to determine where the vertical bars were located. Knowing the location of the signals, a time delay could be calculated, which corresponded to the unwanted source, and this could be used spectrally to remove the unwanted speaker.

The major problem in the use of dual sensor speech enhancement, is the distance between the microphones that is required. As the dual case is a very simple microphone array it can be seen that as the number of sensors decreases, the space between them that is required to effectively steer a beam increases. This could be seen practically from the examples of [23] and [4], where the difference in sensor position ranged from 4 cm in [23], to 75 cm in [4]. Using the techniques which were shown here, it would be very difficult to apply a two sensor approach,

as the distance required between the reference and primary microphone, would be far larger than the dimensions of a typical mobile handset. It can be seen over the last 10 years, how the mobile handset has changed from a briefcase sized device with limited battery time, to a device that can be smaller than the palm of the person using it, with a longer battery life, and the ability for internet access. Given this reduction in size, it is impractical to believe that users would wish to use a handset which is large enough to allow the spacing required for these two microphone and adaptive filtering/beamforming approaches.

A straightforward ANC or beamforming approach is not the best way forward for mobile phone speech enhancement, and for this reason , the decision was made to look at spectral domain based speech enhancement methods, to see if they could overcome the issues of convergence, operation in a non-stationary environment, and the placement of the sensors.

2.4 Spectral Techniques

The influence of a corrupting signal on speech is a twofold process, there is content of the corruption, and the power of the corruption. In temporal techniques such as ANC and beamforming, the entire corrupted signal is operated on. The time taken for convergence of adaptive filters has already been seen as a drawback of some temporal noise reduction designs. This leads to the notion, that perhaps an advantage in the noise reduction performance could be achieved, if only the most relevant portion of signal is altered. Transformation of the signal into the frequency domain splits the signal into real and imaginary components, which can be combined to give magnitude and phase. The difference between these two portions are:

1. The magnitude spectrum represents the relative amplitudes of the frequency components.
2. The phase contains the information, which constructs the content of the signal.

This points to the fact that, using an arbitrary magnitude spectrum speech can be reconstructed using the original phase information. While the speech may not sound natural, due to the altered distribution of the magnitude, it would be possible to understand the content of the speech. On the other hand, if the original magnitude spectrum is combined with an arbitrary phase component, then the content of the resultant speech would sound nothing like the original. This is an idea that has been examined for a long time in image processing leading to the concepts of phase only synthesis [29].

For speech enhancement, the important aspect is to determine whether to operate on the magnitude or phase portion. Conceptually, it seems that operation on the magnitude would be the way to go, as reducing the interfering signal to a level at which it could not be heard, is essentially speech enhancement. It does not matter if the phase content is still present in the signal, reducing the magnitude will make it less audible. This does not mean that phase is unimportant in the overall speech, but in speech enhancement it is not generally operated on. In [8] the importance of phase is discussed as:

> "...speech quality depends on a good representation of the short time spectral magnitude, whereas phase is relatively unimportant"

Some work was presented on the importance of phase for speech enhancement in [30], where it was shown that the SNR of enhanced speech does not vary much at all, if the accuracy of the phase was altered slightly. It stated:

> "... we conclude that an effort to more accurately estimate the phase from noisy speech is unwarranted in the context of speech enhancement, if the estimate is used to reconstruct a signal with an independently estimated magnitude.."

2.4.1 RASTA processing

The conclusions of [30], showed that the greatest results for speech enhancement could be obtained through operation on the magnitude portion of the frequency spectrum. A technique of interest was shown in [31] and [32], which was specifically for the condition of cellular phones called the Relative Spectral (RASTA) processing. The outline of this can be seen in figure 2.6.

The input speech was transformed into the frequency domain by the FFT, and then translated into cubic root power. This step was performed as RASTA processing involveed non-linear filtering of the short time power spectrum of the signal. The filtering in [31] was performed using FIR Wiener filters with each filter designed to map a time window of a specific frequency of the noisy speech spectrum, to a single estimate of the short time magnitude frequency of the clean speech. Each filter covered an 8ms portion of the signal, and the frequency response of the filters was initially designed using speech recorded over a cellular phone line. The entire

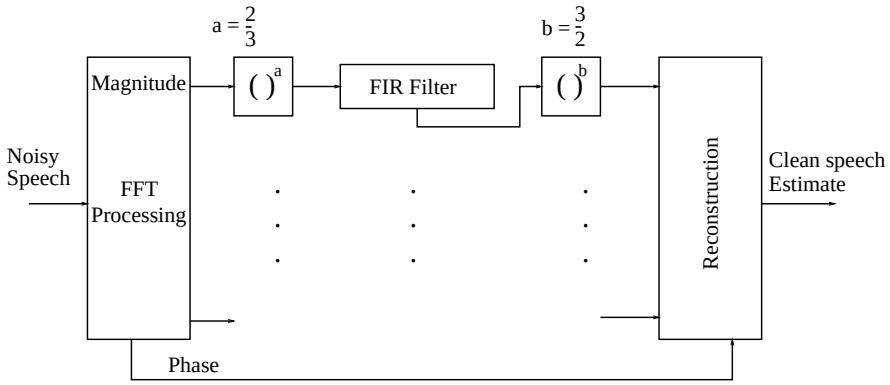


Figure 2.6: *Block diagram of the RASTA process*

telephone bandwidth was covered by 129 filters, each with their own frequency response.

The highest gain filters in the RASTA process lay in the $300-2300$ Hz range, and were designed to be band pass filters with modulation frequencies of $6-8$ Hz. The filters for the $150-250$, $2700-4000$ Hz frequency range were low gain low pass filters, and those for the $0-100$ were high gain filters, which exhibited a very flat frequency response. Experiments performed with noisy cellular speech were claimed to bring about a noticeable improvement in the quality of the enhanced speech, with the noise being less noticeable. Results of RASTA processing in additive and convolutional noise can be found in [33], where it was claimed that operation on noisy speech by RASTA processing should alleviate spectral components, due to additive noise, although it was pointed out, that a degree of post processing was required to deal with the negative values which were introduced into the power spectrum.

A point which was noted in this paper, was that noise components which were additive and uncorrelated in the frequency domain, cannot be effectively removed by RASTA processing, as these components become signal dependent after the logarithmic operation of the RASTA processing. This meant, that RASTA processing was not appropriate for speech which had significant additive noise, a condition which may be implemented in the examination of the intended mobile phone application.

2.4.2 Cepstral techniques

It can be seen, that the framework of a dual microphone technique can be used, if the system is enhanced by the use of signal processing techniques, to overcome the lack of physical sensors. Another area of research has centred around the use of cepstral coefficients, which were defined

in [34] as the inverse Fourier Transform of the logarithm of the power spectrum of the signal:

$$c(n) = FFT^{-1}[\log(X(m))^2] \tag{2.3}$$

where $c(n)$, $(X(m))^2$ are the cepstral coefficient and the power spectrum of the signal respectively.

One of the main uses of cepstral coefficients, is in the detection of speech/pause sections in signals [35][36]. In these the coefficients of the speech and the background are calculated. By examination of the difference between these two sets of coefficients (the cepstral difference), it is possible to determine where sections of speech begin, and construct a speech detection system. Setting a threshold allows the detection system to determine at what point the size of the cepstral difference denotes that speech is present. An expansion of this examines whether it is possible to use cepstral speech/pause detection systems for the enhancement of noisy speech. In [8], it was pointed out that one of the attractive aspects of cepstrum techniques was the ability to examine signals, in which there is no way of knowing how the parts have been combined, which could be useful for mobile phone use.

Work on speech recognition in noise, using cepstral time based features was presented in [37] which was used to aid noise reduction, by either spectral subtraction or state based time varying Wiener filtering. After calculation, the cepstral coefficients were used to provide information on the time variations of the signal. In the case of noise reduction by Wiener filtering, this allowed for the implementation of a more accurate filter. This technique was interesting, as the speech enhancement section used the noisy speech and an estimate of the noise in the spectral domain. The cepstral features were not used directly to reduce the noise in the signal, but to recognise when speech was present, to improve the enhancement process by the prevention of removal of any speech sections.

The majority of cepstral work for speech centers around dealing with reverberations, estimating speech/pause areas or pitch estimation. It seems that the area of enhancing speech corrupted with additive noise is not an area that cepstral techniques are easily able to deal with. This suggests, that cepstral features may not be as robust when applied to the area of noise reduction, as opposed to echo control. The work presented in [38], examined how Gaussian noise affected short term cepstral analysis of speech. It was shown that cepstral based speaker identification

performance degraded severely, unless there was a clean speech model present with which to train the recogniser. This was to be expected, as the availability of clean speech for training is an idealised case. It was proposed that the effects of Gaussian noise on the logarithmic power spectrum was complex (as shown in equation (2.4)), and that by transforming this into the cepstral domain could further compound this.

$$\log(P_Y(m)) = \log(P_S(m) + P_N(m)) \tag{2.4}$$

where $P_Y(m),\ P_S(m),\ P_N(m)$ are the power spectrum of the noisy speech, clean speech, and the noise respectively.

The effects of Gaussian noise was shown to be a change in the distribution of the signal, and an alteration of the mean and variance. As the noise component was increased the distribution of the noisy speech cepstral components shifted, to mimic that of the corrupting noise. It was suggested that before any cepstral calculations took place, a compensation factor was required to remove the degradation in the distribution, mean, and variance. The drawback of such a factor in non-stationary case was that the compensation would require constant updating to maintain accuracy of the effects of the noise. This introduced delay into the system, and could prove difficult to achieve when there was no access to the clean speech, with which to compare the compensated result.

It was clear, that the use of techniques such as RASTA and cepstrum processing were not applicable to the specific problem of reducing non-stationary noise from cellular conversations. It had been shown that through the use of signal processing techniques, it was possible to compensate for the relatively low number of sensors which were used, rather than resorting to a cumbersome microphone array. One technique of particular interest was presented in [37], which is spectral subtraction.

2.5 Spectral Subtraction

2.5.1 The Work of Boll

Spectral subtraction is a technique which was first studied comprehensively by Boll [6], although much work had already been published near that time [39], [40]. The technique operates in the frequency domain, and is based on the direct estimation of the short time magnitude of the clean speech. It has already been seen, that operation on the magnitude portion only is a viable method of speech enhancement. Some assumptions are made for spectral subtraction, which mirrored the mobile phone problem to be investigated.

1. The background noise is acoustically or digitally added to the speech.
2. The added noise is uncorrelated to the speech.

It was also assumed in [6] and [8], that the additive noise was short term stationary, something which would not be assumed for the conditions being investigated. This assumption was only critical when one sensor is being used to detect both speech and noise. In [6], it was stated that any change in the noise leading to a new stationary state required a space of 300 ms, in order to ensure the algorithm had captured a good estimate before any speech occurs. If a secondary sensor was used, then this criterion could be relaxed and the secondary sensor could track the noise conditions at all times, even during speech segments.

In [6] the additive noise model was denoted by:

$$y(n) = s(n) + n^{'}(n) \tag{2.5}$$

where $y(n),\ s(n),\ n^{'}(n)$ are the noisy speech the clean speech, and the background noise respectively. The outline of spectral subtraction is shown in figure 2.7

The time domain signals were multiplied with half overlapping Hamming windows, and the resultant data was Fourier transformed as shown in figure 2.7. From this Fourier representation, the magnitude and phase of the signals were calculated, and the phase of the background noise was discarded, with reconstruction of the signal occurring using only the corrupted speech phase. In order to perform the speech enhancement, it can be seen in figure 2.7 that an estimate

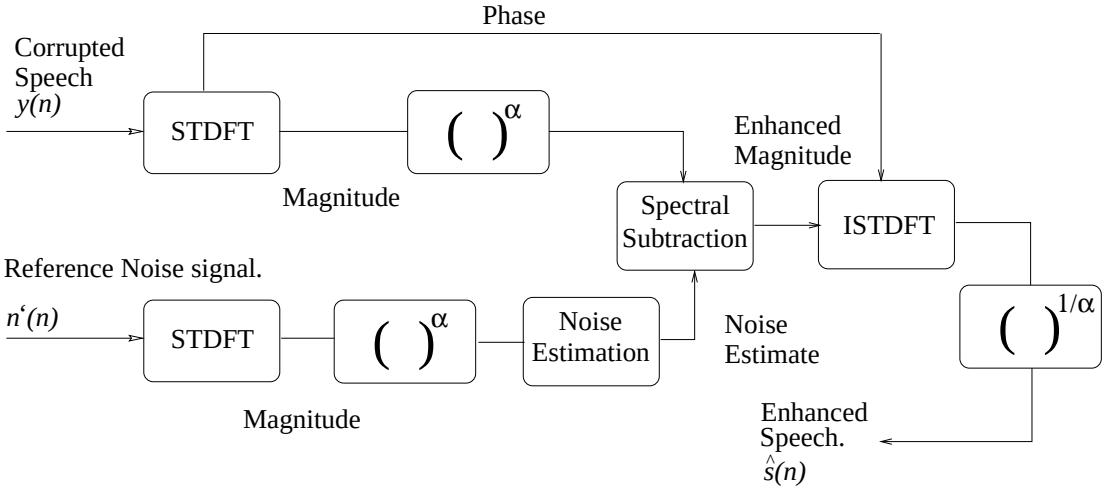


Figure 2.7: *Block diagram of the spectral subtraction process*

of the background noise must be produced, and this was then subtracted from the corrupted speech:

$$\widehat{S}(m) = (|Y(m)|^{\alpha} - \beta|\widehat{N}(m)|^{\alpha})^{\frac{1}{\alpha}} \tag{2.6}$$

where $\widehat{S}(m),\ |Y(m)|,\ \widehat{|}N(m)|,\ \alpha,\ \beta,\ m$ are the estimated clean speech spectrum, corrupted speech magnitude, estimated noise magnitude, power and oversubtraction factors, and the frequency index respectively.

Equation 2.6 shows the generalised form of spectral subtraction, which by alteration of the parameters α and β, can produce different versions of the subtraction process. An overview of the effects caused by this can be found in [41]. The estimation of the background noise is an important part of spectral subtraction, which shall be discussed in section 2.7. One thing that can be assumed about the noise estimate, is that it will be non-ideal, which will lead to a spectral error shown in [6] as:

$$\epsilon(m) = \widehat{S}(m) - S(m) \tag{2.7}$$

where $\epsilon(m),\ S(m)$ are the spectral error, and the clean speech spectrum respectively.

In order for the quality of the enhanced speech to be high, it is desirable to reduce this spectral error as much as possible. In [6], four different techniques were suggested to achieve this:

1. Magnitude averaging
2. Half wave rectification
3. Residual noise reduction
4. Additional attenuation during speech pauses

Further explanation of these can be found in [6] and [8]. After the estimated clean speech has been produced, the frequency domain representation is transformed back into the time domain.

2.5.2 Applications of spectral subtraction

As the technique of spectral subtraction is conceptually easy to understand it has been applied to a wide variety of enhancement situations. Due to this, it is of particular interest for mobile phone enhancement, because of the potential to use either one or two sensor systems. As this is a spectral technique, the signal being enhanced need not be speech, but any audio signal. This is a popular area for enhancement techniques, with many products being produced for consumer audio, specifically to remove the effects of scratches and artefacts produced on vinyl, for recording onto CD or minidisc.

The subtraction method in [42] attenuated sections of music, which were strongly corrupted by noise, the main criterion being the reduction of spectral distortion, rather than the reduction of the noise. Once again, the transformation technique was the Fourier transform, which was used to produce the power spectrum of the signal. The noise reduction performed was the result of multiplying a gain (< 1) to the power spectrum of the noisy speech. This gain was applied to each frequency component in the frame according to a specific *"noise suppression rule"* given by:

$$G(i:m) = \left(1 - \frac{1}{Q(i:m)}\right)^{\frac{1}{2}} \tag{2.8}$$

where $G(i:m)$, $Q(i:m)$, i are the noise suppression gain, the relative signal level, and the frame index respectively.

The relative signal level is the relation between the noisy speech spectrum, and the estimated noise power spectrum. This controlled the amount of gain which was applied to the spectrum.

One spectral subtraction method which is close to the described mobile phone scenario was presented in [43], where it was proposed that the most important aspect for speech enhancement was to make the speech intelligible. While this is undoubtedly true for many applications, (Army field radio, air traffic control communications), a consumer device such as the mobile phone also requires that the output from the algorithm be pleasing to the ear. If this is not the case, then it would be hard to convince consumers of the benefit of such a feature in a handset. It was discussed that in monoaural hearing (e.g. telephone handset), the listener loses the directional cues used in normal binaural hearing to aid enhancement by the auditory system. This meant that for monoaural hearing, the desired signal must have greater power than the corruption, in order for the listener to be able to hear and understand it.

The work in [43] also showed how a dual microphone spectral subtraction technique could give greater enhancement, due to the ability to identify not only the spectral energy distribution of the noise, but its time varying nature as well. Single sensor methods only examine the energy distribution, as the noise can only be updated during periods of non-speech activity, between which the noise is assumed stationary. In [43] was claimed that for negative SNR circumstances, spectral subtraction would perform well with general noise (stationary or non-stationary), or interfering speakers. The drawback of the use of spectral subtraction in such conditions, was the potential to introduce musical noise (which will be introduced in section 2.7) into the processed signal.

An overview of some of the subtraction algorithms produced for the enhancement of additive noise corrupted speech was given in [40], where the dangers of making too many assumptions about the speech signal were shown. It was claimed that, as more assumptions were made about the speech signal, the enhancement system would become more sensitive to any derivation from the assumptions. This is also true about the assumptions which are made about the corrupting noise.

This showed, that in order to maintain a robust spectral subtraction based algorithm, there should be a minimal number of assumptions.

The ability to enhance signals in either stationary or non-stationary noise is desirable for the enhancement of mobile phone speech. One technique which is designed to operate in these conditions was presented in [44], where the speech was enhanced with an algorithm based on standard spectral subtraction. The section which allows the enhancement in non-stationary conditions was the noise estimation algorithm, which runs continually during both speech and pauses, producing new estimates of the noise. The spectral subtraction itself conformed to the layout of figure 2.7. The noise estimation algorithm of this technique will be examined in section 2.7.

It could be seen, that the spectral subtraction technique was a promising method for the reduction of non-stationary noise. It had the advantage of being easy to compute but with the flexibility to add other signal processing elements to improve the performance. By application of accurate noise estimation techniques and dual sensors, it should be possible to enhance speech corrupted by stationary noise, non-stationary noise, and competing speakers. The only drawback to the operation of this technique, was the possibility of introducing musical noise into the processed speech. It is possible that with the use of extra signal processing techniques, this problem could be overcome.

2.5.3 Choice of frequency transform

The only choice to be made before any work was performed by the spectral subtraction process, was which transformation should be used to translate from the time to the frequency domain. There were three main transforms which could be chosen, the discrete Fourier transform (DFT), the discrete cosine transform (DCT), and the wavelet transform. Previously the FFT was described as a transformation method, For discrete digital data, the DFT is commonly computed using the FFT.

The DFT is based on the Fourier series. This assumes that a signal which is continuous and periodic can be decomposed into a sum of constituent sinusoids of varying amplitudes and phases.

$$S(m) = \frac{1}{N} \sum_{n=0}^{N-1} s(n) e^{-j\frac{2\pi}{N}mn} \tag{2.9}$$

$$s(n) = \sum_{n=0}^{N-1} S(m) e^{j\frac{2\pi}{N}mn} \tag{2.10}$$

where $s(n),\ S(m),\ m,\ n,\ N$ are the time domain speech, the frequency domain spectrum, the harmonic number, sample number, and number of samples per period respectively.

As part of the Fourier transform, family the DFT is one of the classical methods of transformation, from the time to frequency domain. When transforming the signal using the DFT, it results in a portion of the signal represented as cosine waves (the real part), and a portion represented as sine waves (the imaginary part). The DFT is useful, in that from the real and imaginary aspects of the signal, the magnitude and phase of the speech signal can be calculated.

$$|S(m)| = \left(S(m)^2_{\mathrm{Re}} + S(m)^2_{\mathrm{Imag}}\right)^{\frac{1}{2}} \tag{2.11}$$

$$\Theta_S(m) = \arctan(\frac{S(m)_{\mathrm{Imag}}}{S(m)_{\mathrm{Re}}}) \tag{2.12}$$

where $|S(m)|,\ S(m)_{\mathrm{Re}},\ S(m)_{\mathrm{Imag}},\ \Theta_S(m)$ are the magnitude spectrum, real portion of the signal, imaginary portion of the signal, and the signal phase spectrum respectively.

As most operations on speech processing work on the magnitude portion of the speech, this transformation technique is useful.

Whereas the DFT uses both sine and cosine waves to represent the signal which has been transformed, the DCT uses only cosine waves, meaning that the function is only real valued. Due to this, it is possible that the DCT will be faster than the DFT, although as the DCT is computed using FFTs this is not always true.

The DCT and the inverse DCT are described by:

$$S(m) = \sqrt{\frac{2}{N}} E_m \sum_{n=1}^{N-1} s(n) \cos\left(\frac{(2n+1)m\pi}{2N}\right) \quad \text{where } m = 0, 1, \ldots N-1 \tag{2.13}$$

$$s(n) = \sqrt{\frac{2}{N}} \sum_{m=1}^{N-1} E_m S(m) \cos\left(\frac{(2n+1)m\pi}{2N}\right) \quad \text{where } n = 0, 1, ...N-1 \tag{2.14}$$

$$E_m = \begin{cases} \frac{1}{\sqrt{2}} & : \quad m = 0 \\ 1 & : \quad \text{otherwise} \end{cases}$$

$s(n),\ S(m),\ m,\ n,\ N$ are the time domain signal, the frequency domain spectrum, the harmonic number, sample number, and number of samples per period respectively.

The possible drawback of the DCT was given in [45], where it was thought that as the DCT performs signal analysis with purely cosines, it was not as useful for the analysis of purely sinusoidal waveforms. This is not to say that the DCT is unable to examine sinusoidal signals, it requires longer to produce the requisite sum of cosines to produce a sinusoid.

Due to this, the DCT is not applied in speech enhancement systems, and its main application is in the area of signal compression.

Wavelet transforms differ from the previous DFT and DCT techniques, in that while the DFT and DCT work with fixed resolution, the wavelet transform alters resolution according to frequency as shown in figure 2.8.

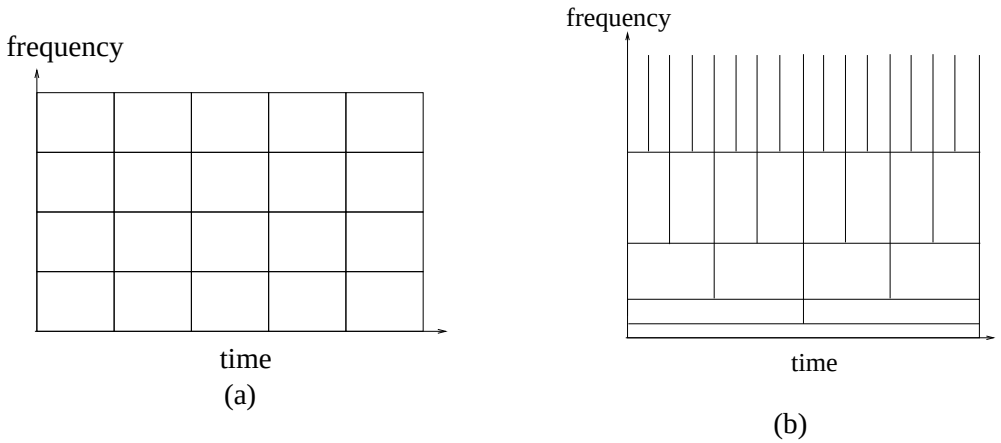


Figure 2.8: *Coverage of the time-frequency plane for the (a) Short time DFT and DCT (b) Wavelet Transform*

It can be seen in figure 2.8, that whereas the DCT and DFT time frequency resolution stays fixed, the wavelet resolution varies in both time and frequency, leading to a mulitresolution

analysis. The change in frequency resolution is always proportional to the center frequency of the analysis region:

$$\frac{\Delta f}{f} = c \tag{2.15}$$

$\Delta f,\ f,\ c$ are the change in frequency, the center frequency, and a constant which determines the ration between the two. This change in time and frequency resolution can be seen in figure 2.8(b), where it is seen that time resolution becomes good at higher frequencies, with frequency resolution better at lower frequencies. In [46] it was stated that:

> "The generalisation of the concept of changing resolution at different frequencies is obtained with so called "wavelet packets", where arbitrary time-frequency resolutions are chosen depending on the signal."

The WT does this by modelling all sections of the signal, by scaled versions of a prototype or kernel.

$$h_q(n) = \frac{1}{\sqrt{a}} h(\frac{n}{a}) \tag{2.16}$$

$a,\ n$ are the scaling factor and the time index. Through combinations of the scaled versions of this mother wavelet, the Continuous Wavelet Transform (CWT) is:

$$CWT_x(\tau, a) = \frac{1}{\sqrt{a}} \int x(n) h^*(\frac{n-\tau}{a}) dn \tag{2.17}$$

τ is the change in time.

Wavelet analysis is a new technique, which has found applications in computer and human vision, fingerprint compression [47], and the denoising of data [48]. It was of particular interest, because of the theoretical fitting of the wavelet to non-stationary signal analysis, which would be of use for the mobile phone case being examined.

One of the largest problems with wavelet analysis though, is the choice of the mother wavelet.

One of the most interesting questions is, how to choose an optimal wavelet to form the basis function for a given application.

Having examined the transformation methods which could be applied to the speech scenario, the following observations could be made:

1. The DFT is a well known algorithm, which can be quickly calculated via the FFT. It has the drawback of analysing signals in a fixed resolution, but does provide the magnitude and phase elements required for speech enhancement.

2. The DCT can be faster than the DFT, and can also be calculated via the FFT. It also analyses in fixed resolution, but has the problem of requiring longer to operate on sinusoidal signals.

3. The WT gives a multiresolution analysis of the signal, and has the possibly of operating on non-stationary signals. The drawback is the difficulty in selecting a suitable mother wavelet for the analysis.

At this point it was decided to use the DFT as the method of transforming the speech information from the time, to the frequency domain. While it may not be able to perform the multiresolution analysis of the WT, it provides the frequency domain in a manner which is suitable for speech enhancement. If it were possible to apply a form of non-linear analysis to the frequency domain signal after transformation, then the advantages given by the WT would be realised.

2.6 The human auditory system

One of the problems of spectral subtraction for speech analysis, is that it takes no account of the way the human ear perceives sound. Any speech enhancement technique which ignores perceptual effects of the enhancement will be unable to shape the enhancement, so that the speech sounds natural to the listener. It is also true, that when humans hear sounds, elements of the sound field can be covered by louder signals, which occur concurrently in time or frequency. This is called masking. The human auditory system is very good at enhancing a signal of interest in noise, and if the reason for this could be understood, it may be possible to implement such a technique into spectral subtraction, increasing the potential performance. It is known

that part of the reason for this is the binaural hearing offered by the human ears, and neural processing. Despite this, there is still a degree of mechanical processing performed in the auditory system.

Perceptual criterion have been used in a variety of ways for signal processing. One area of work is the use of perceptually modelled filters for signal enhancement, with speech or general audio [49, 50]. The idea behind this, was to design filters which would enhance the signal in a manner similar to the auditory system. Through doing this, it was hoped that any noise which was present in the signal would be shaped in such a way, that it became perceptually unimportant to the listener.

One of the largest areas of perceptual study is in the area of signal coding, where it is hoped to reduce the data rate and compress the transmitted signal [51–60]. This compression is performed by using the auditory modelling to establish which portions of the signal are audible, and only operate on these, all portions deemed inaudible are removed. This methodology is of course the basis for many of the audio coding standards used for digital audio in cinema (Dolby Digital, DTS, MPEG) or in mini-disc (ATRAC) [61–65]. This allows the signal to be reconstructed digitally in a manner where the majority of the signal quality is still apparent, but there has been a significant reduction in the data rate.

A new area of research is the application of auditory criterion to the enhancement of speech signals. By applying the ability of the ear to pick out signals above background noise, speech enhancement can be significantly improved with little extra processing [66–76]. The key to the use of auditory aspects for speech enhancement, is to determine which portions of the auditory system perform the noise reduction.

There is a great deal of work on the mechanics of the auditory process of humans, how it operates, and whether it is possible to simulate this operation for digital signal processing [77–92]. figure 2.9 shows the mechanical processes which make up the auditory process, with the major areas highlighted. The basic functions of the main sections of the auditory process shall now be outlined. This is not intended to be a comprehensive discussion of the auditory process, and more information on this can be found in [2, 3, 80, 93]

2.6.1 Outer Ear

The outer ear consists of the pinna and the ear canal, as can be seen in figure 2.9. The pinna is primarily a filtering device which operates on air vibrations and its main function is in the localisation of sounds [2]. This is a process which is most visible in animals who are able to rotate their ears towards the direction of sounds of interest, such as cats and dogs.

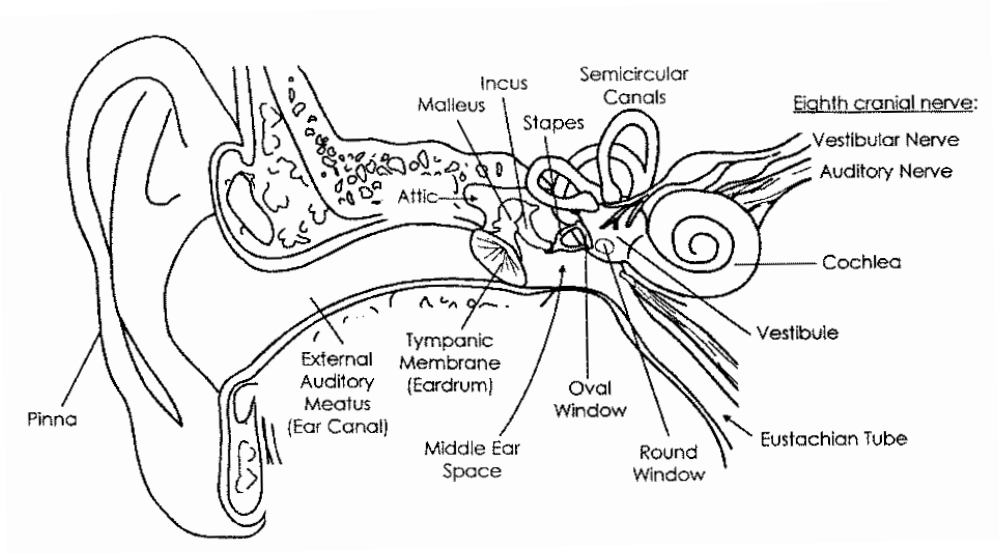


Figure 2.9: *The human auditory process as shown in [2]*

The pinna is able to locate sounds, even if the signal is only applied monoaurally [3, 93] which is important, as it shows that not only the interarrival times between ears is used for the localisation of sounds. Without the pinnae, the ability to localise sounds in terms of the vertical plane would be lost, although the interarrival time would mean that it would still be possible to localise in the horizontal plane.

The ear canal runs from the pinna to the ear drum, narrowing as it moves deeper into the head. Its main function is to provide a channel for the propagation of the sound waves to the ear drum. As the ear canal is a tube which is sealed at one end, and open at the other, any sound which enters it cause resonances. It has been found that the ear canal has a resonant frequency of approximately 3800 Hz, and this affects the frequency distribution of any signal which passes through the ear canal. This means that the ear canal has a transfer function, which can be seen in figure 2.10

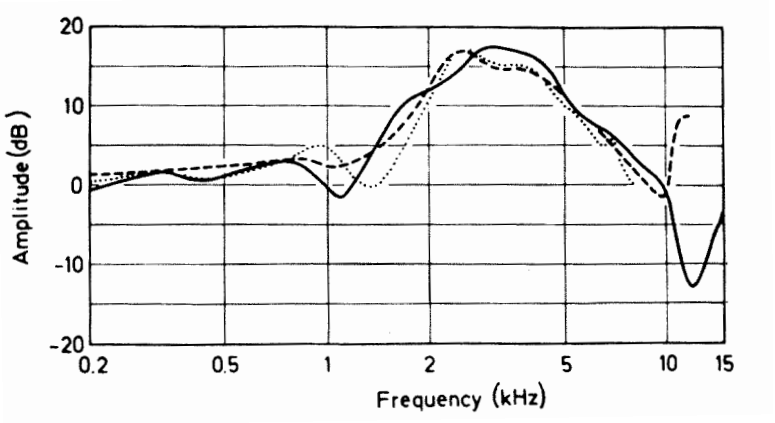


Figure 2.10: *The transfer function of the ear canal as shown in [2] by three different studies*

2.6.2 The middle ear

The middle ear consists of the eardrum, malleus, incus and stapes, the latter three being small bones which connect the movement of the eardrum to the cochlea.

The ear drum is the first part of the middle ear, and has the role of converting the information present in the air pressure waves into the fluid vibrations in the middle ear. It can be seen in figure 2.11, that the eardrum is a diaphragm, and it acts to match the impedance of the outside air to that of the cochlear fluid allowing all of the information to pass without loss. This step is important as the ratio between the impedances of the cochlear fluid and air is $4000 : 1$ [2]. Without the middle ear section, only approximately 0.1% of the airborne energy would be passed to the cochlea.

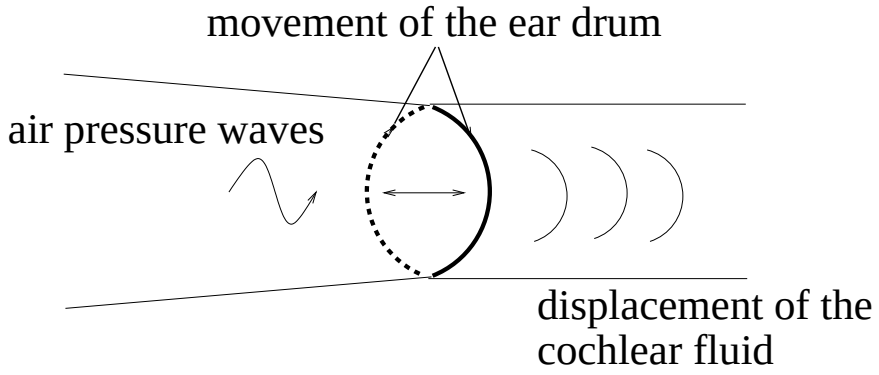


Figure 2.11: *A simplification of the operation of the ear drum.*

The malleus, incus and stapes are collectively called the ossicles, and these act as a lever system in tandem with the eardrum, to regulate the range of pressures which can occur from differing

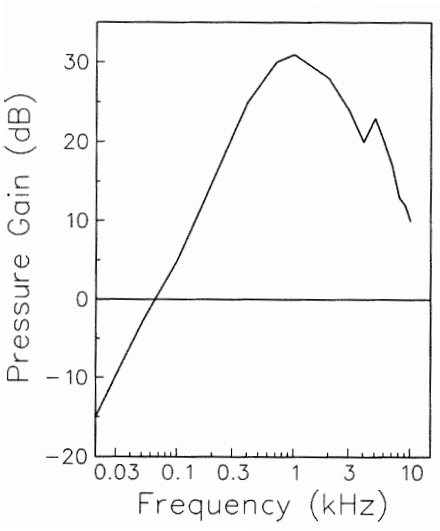


Figure 2.12: *The transfer function of the middle ear as shown in [3]*

volumes. This depends on the frequency of the signal, and can be seen in figure 2.12.

It can be seen, that so far the outer and middle ear can be seen as conduction and filtering devices, which have no part in noise reduction. The only mechanical section left which could provide this, is in the inner ear.

2.6.3 The inner ear

From figure 2.13, it can be seen that the inner ear consists of the semicircular canals, the vestibule, and the cochlea. The semicircular canals along with the vestibule, under normal circumstances, affect the sense of balance and not hearing. The stapes connects to the vestibule, to provide a path for the transmission of information from the eardrum to the cochlea.

The inner ear is filled with a fluid called perilymph, which provides the correct chemical environment to transfer the vibrations from the middle ear to the sensory units in the cochlea. Within the vestibule are organs which allow the body to deal with the effects of linear acceleration and gravity, whereas the semicircular canal ampullae deal with rotational acceleration. While all of these deal with non-auditory facets, the primary auditory organ in the inner ear is the cochlea, which is a small shell shaped organ, as can be seen in figure 2.13. The spiralled cochlea is typically only 35mm long, and inside there are three fluid filled ducts, two of which are separated by the basilar membrane.

The movement of the auditory information from the stapes is conducted by the displacement

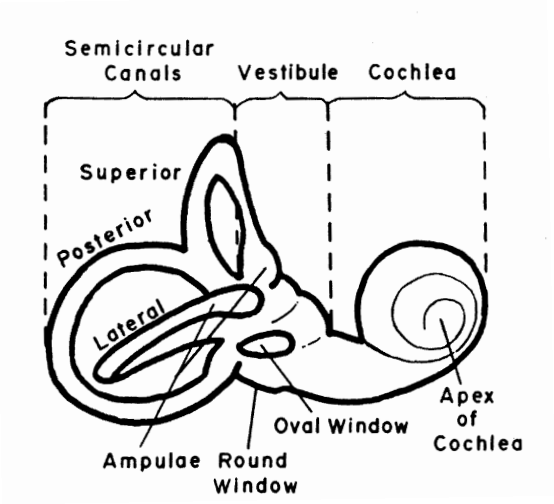


Figure 2.13: *The parts of the inner ear as shown in [2]*

of the perilymph fluid over the basilar membrane (BM). This displacement causes the BM to respond mechanically, mirroring the displacement of the fluid as shown in figure 2.14.

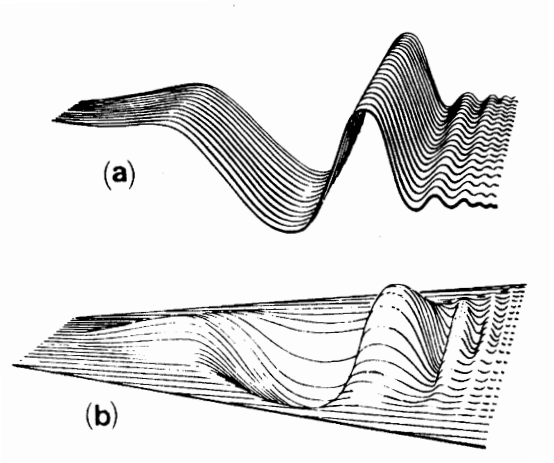


Figure 2.14: *Vibration of the BM if it were (a) like an unconstrained ribbon (b) constrained as in the cochlea as shown in [2]*

It is this displacement of the BM which essentially gives the ability to hear. Each sound that is picked up by the BM causes it to vibrate in a specific region, corresponding to a frequency. The frequency resolution of the BM is logarithmic above 500 Hz, and linear below that value. In signal processing, this is modelled by the use of critical bands, which are said to be the separation between two frequencies, in order for them to be heard individually.

As the BM is continuous and constrained, an excitation in one part of the membrane will cause vibrations in other sections. If this displacement is bigger than one caused by another desired

sound in the same portion of the BM, then it will be swamped or 'masked" by the louder concurrent sound. This is the concept of frequency masking, where sounds of the same frequency occur concurrently in time. The louder signal will cause a larger vibration of the basilar membrane, causing other sounds to be affected by the sympathetic vibrations which occur in the membrane. There has been a great deal of work into modelling how this action occurs in the cochlea [59, 81–83, 92], and by using this ability, it is hoped that it would be possible to operate only on those portions of the noise which are not masked by speech. This leads to a reduction in computational time for the enhancement, as only those background points which are not masked are operated on. There is also a secondary effect of applying this technique, as it processes the signal in a manner more natural to the ear, meaning that any noise which is present in the signal will be shaped into a form which is consistent with the way the ear would deal with it.

The point of using perceptual techniques, is to enhance the spectral subtraction process as it currently stands. It is important to remember, that the ear does do this process but in the monoaural listening environment it finds this a far harder task to perform. Processing the speech in this manner, before it is coded and transmitted, allows for a more perceptually accurate signal to be presented to the listener.

2.7 The importance of a good noise estimate

For the proposed speech enhancement technique, only the corrupted speech and the background noise will be present. In order to be able to obtain good performance from the enhancement algorithm, it is important that the estimate of the background environment being presented is as accurate as possible. As spectral subtraction relies on this estimate to remove the corruption from the speech, any error in the estimate leads to an error in the estimation of the enhanced speech. This error can manifest itself three ways:

1. The presence of a residual background noise signal
2. The appearance of random tonal artefacts called “musical noise”
3. A loss of speech data

This can be seen in figure 2.15, where a section of clean female speech is presented, along with the noise corrupted case, and the enhancement after subtraction of an inaccurate noise estimate. In figure 2.15 (c), it is clear that not only has a substantial part of the speech been removed, but there is an amount of noise still present in the signal. It should be noted that this noise is not of the form seen in the corruption of figure 2.15 (b), and is more discrete in the frequency domain. This is the so called "musical noise" phenomenon, which manifests as a water-like warbling sound in the processed speech. The spectrograms in figure 2.15 serve to show the importance

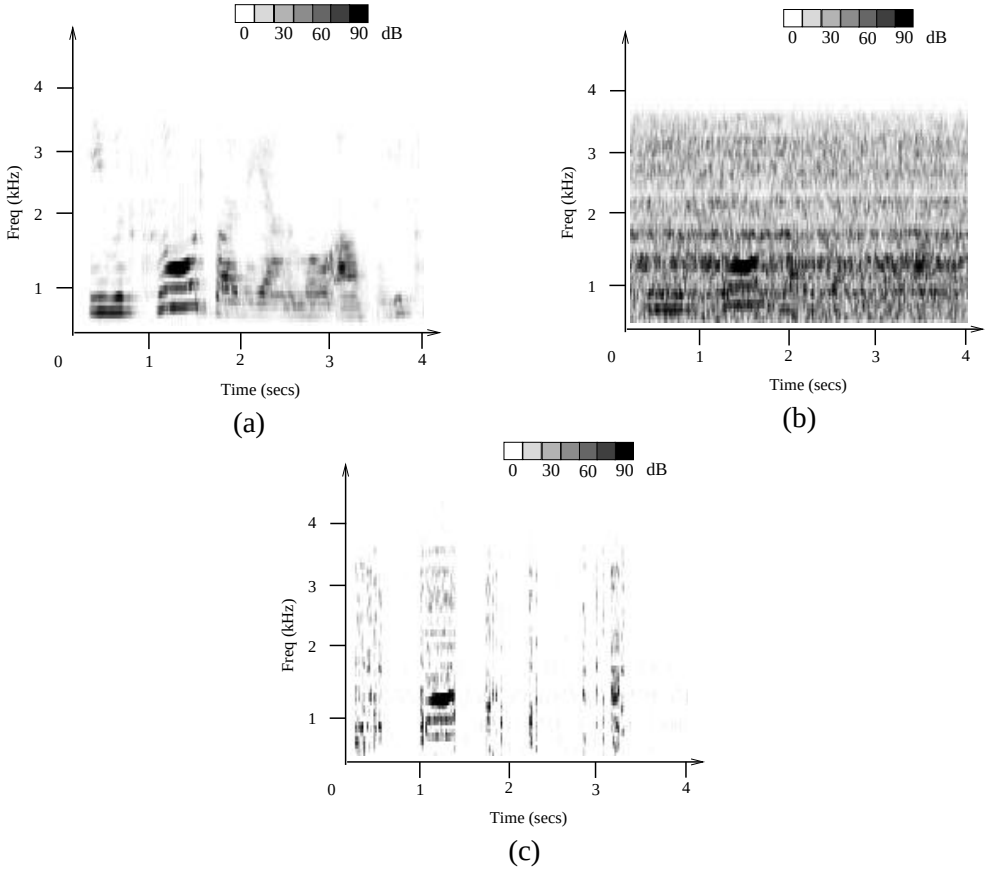


Figure 2.15: *Spectrograms of (a) Clean Female Speech (b) Speech degraded with pink noise (c) Enhancement after subtraction of an inaccurate noise estimate*

of applying perceptual techniques to the spectral subtraction process.

2.8 Summary

In this Chapter a background into some of the techniques used for speech enhancement was discussed, with the methods of adaptive filtering, and multiple microphone beamforming being

examined. It was seen, that the typical time it took for the filter coefficients to adapt to a new noise environment was too long to allow this technique to work well in non-stationary conditions. Beamforming techniques presented some promising results, but they suffered from the need to use multiple sensors which was restrictive for mobile phone implementation.

A couple of spectral techniques were examined which looked to remove the noise in the frequency domain, as opposed to temporally. RASTA and cepstral enhancement algorithms were examined, and by using these it is possible to use a two instead of multisensor based algorithm. It was shown how RASTA processing work was designed to work with convolutional rather than additive noise, which was a severe drawback for use in the proposed mobile phone case. Cepstral based algorithms suffered from the problem that cepstral coefficients are severely degraded by noise causing the distribution, mean, and variance of the signal to be affected detrimentally. In order to compensate for this, a specially calculated compensation factor would need to be applied, which would prove difficult, without access to the original clean speech to compare the compensated signal with.

Spectral subtraction was introduced, as a technique which is widely used in speech and audio processing and some applications, such as the restoration of old musical recordings were examined. The drawback of the possible introduction of musical noise into speech was discussed, although it was felt that it was possible to overcome this, making spectral subtraction a valid methodology. In order to translate the signal to the frequency domain, a suitable transformation method was required. An examination was made of the DFT, DCT and WT, leading to the choice of the DFT.

The drawback of many speech enhancement systems, is that no account is taken of how the human auditory system will perceive the output. If the system could perform the enhancement and produce an output which would be in a form the ear would find pleasing, then this is a big step in improving the enhancement process. To this end, the human auditory process was briefly examined, with regard to their importance in noise reduction. The main section for this was seen to be the cochlea, and the concept of frequency masking was introduced as a method for humans to process signals and pick out relevant information.

Chapter 3
Speech Enhancement by Spectral Subtraction

In this Chapter, the basis for the spectral subtraction (SS) algorithm will be examined. The transformation required from the time to frequency domain shall be presented, along with parameters required to determine the frame size, control of the subtraction, and how to deal with residual values. The auditory analysis shown in Chapter 2 is used in the spectral subtraction, and the effects of this will be shown

3.1 Transformation into the frequency domain.

In Chapter 2, it was established that SS operates in the frequency domain, on the magnitude portion of the presented signal. As the speech data is presented in the time domain to the primary and reference microphones, it requires to be translated into the frequency domain for the SS to be implemented.

In Chapter 2 the various frequency transformation techniques were shown, and the DFT was chosen for this system. The DFT was required, as speech waveforms are generally more complex than a steady sinusoidal signal. Fourier transforms of the entire signal are not appropriate on speech whose properties are time varying. In [34], it was shown that by splitting the discrete signal into small time periods, some of the properties of speech stay fixed within that period. This allows the assumption that during this small time period the speech is stationary, and DFTs can be applied to give the transformation. This splitting up of the signal into small sections called frames allows the DFT to treat the data present in the frame as if it is a short segment of a stationary signal. In order to do this, a choice of window has to be made. From the work which had been studied, it was discovered that only two main window types were used in speech processing. These were the Hamming or Hann window, and the only explanation which could be found for the choice of these was due to their ease of computation and conceptually simple function.

The size of window chosen to apply to the signal has an important bearing on speech processing. In general, as the window size increases, the spectral resolution increases, due to increased information. With speech though, there is the added criterion of choosing a window which is small enough to assume the speech is stationary over its period. If this is ignored, the the DFT cannot be used to obtain the frequency spectrum. The important point is, over which period can speech be determined as stationary? As speech in the long term is a highly non-stationary process, the window period is required to be very small. In [8] for example, it was stated that for windows in the order of 20ms, speech can be assumed to remain stationary.

Speech data was provided from BT for the work which was of bandwidth $0 - 4$kHz, sampled at 8kHz. This meant that a 20ms segment would require 160 samples in the window. The DFT can be implemented efficiently by the FFT, which readily operates on window sizes that are a power of two. There was a choice of a window size either 128 or 256 samples long. These represented periods of 16 or 32ms respectively. It was decided to use the shorter window as it had been reported that while windows in the range of 20ms are useful, it was not recommended to go above 30ms [34].

If the short time frames of speech are multiplied by a window, then due to the shape of the window, values taper off at each end. A degree of overlap is required to ensure that information at the edges of the windowed data are not discarded. Typically a 50% overlap is used, as this results in each sample being effectively multiplied by a unit value, and this was the overlap factor used.

3.2 Spectral subtraction process

Having chosen the window, its size and the overlap factor, the data could be transformed into the frequency domain to begin the enhancement process. In Chapter 2 the concept of spectral subtraction was introduced in equation 2.6. The parameters α and β determined the type of spectral subtraction used, and the percentage of the noise which was to be removed. The effects of these parameters on speech enhancement are described in [41], which gave an overview of how different combinations of α and β values could affect the resultant speech. This meant that a choice had to be made for α and β, and a method chosen to deal with negative values resulting from the subtraction.

3.2.1 Choice of power exponent α

Given equation (2.6), it could be seen that the α value determined the power which was applied to the magnitude spectrum before subtraction. In [41], a number of values are examined ranging from $0.2 - 2.0$, and [39] reports results gained from using α from $0.25 - 2.0$. The special cases of $\alpha = 1$ or $\alpha = 2$ denote magnitude or power spectral subtraction respectively. Through an examination of the RMS magnitude error, spectrograms of the resultant speech, and informal listening tests, it would be possible to determine which value was best suited to providing the optimal balance of noise reduction, and speech quality in terms of listenability and intelligibility. As the algorithm would be operating in a non-stationary environment, and the results in [41] and [39] were conducted in a stationary case, tests were performed on corrupted speech with both magnitude and power SS, to determine which would provide the best results. Cases with non-integer values of α were not used, as these do not make significant difference until near the $\alpha = 1$ or $\alpha = 2$ boundaries. During the experimentation, β was kept at 1, and the data examined consisted of either a male or female speaker, corrupted with pink noise at SNRs of $-10, -5, 0$ and 5dB.

Table 3.2 shows the average RMS percentage magnitude errors which occurred from the SS experimentation. The percentage values are based on the maximum magnitude value of the clean speech signal the algorithm was trying to reproduce.

	Percentage error (%)			
Speech operated and SNR	-10dB	-5dB	0dB	5dB
female $\alpha = 1$	1.716015	1.299322	0.962889	0.705566
female $\alpha = 2$	2.543467	1.625185	1.014425	0.666293
male $\alpha = 1$	3.896944	2.949769	2.178746	1.475190
male $\alpha = 2$	5.485017	3.463228	2.209839	1.3319

Table 3.2: *Average RMS percentage magnitude error results for magnitude and power spectral subtraction.*

It can be seen in Table 3.2 that the RMS errors for magnitude spectral subtraction are consistently lower for both male and female speech, until the SNR reaches 5dB. This may be the point where the benefits of power spectral subtraction begin to operate, as the majority of work performed for speech enhancement has dealt with fairly large SNRs. As the RMS error gives a good guide to the corruption of the speech which occurs during enhancement, this points to magnitude SS being a better choice than power SS. For speech though, purely statistical values

cannot be relied on to give the full results. Examination of the spectrograms of the resultant speech would give a clearer indication of which approach was better.

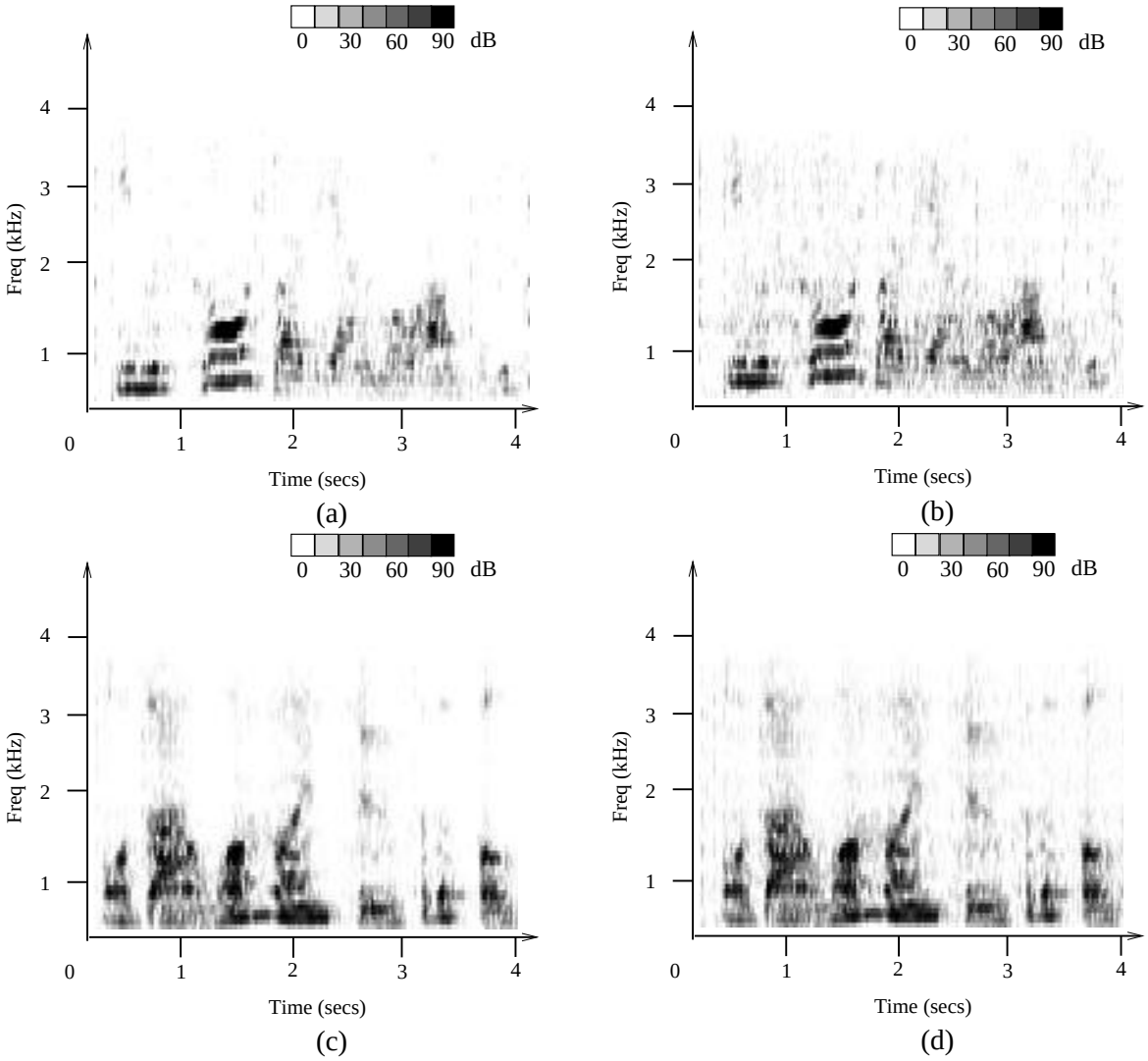


Figure 3.1: *Portion of a Spectrogram (a) Female Speech* $\alpha = 1$ *(b) Female Speech* $\alpha = 2$ *(c) Male Speech* $\alpha = 1$ *(d) Male Speech* $\alpha = 2$

The presented spectrograms are an example of the typical results found by the application of different SS algorithms. Figures 3.1(a) and 3.1(b) represent a portion of female speech *"clawed by the animal after he"* and figures 3.1(c) and 3.1(d) represent male speech *"or dial one nine two for the"*. Comparing figures 3.1(a) and 3.1(b) to 2.15 (a), it can be seen that the $\alpha = 1$ case presents less distortion in the resultant speech, which is most visibly evident in the range $1.5 \rightarrow 3$kHz. In the male case figure 3.1(c), $\alpha = 1$ produces significantly less distortion that the speech produced by the $\alpha = 2$ resultant speech. This was an interesting result, as it showed

the importance of taking more than purely statistical results in speech enhancement. While the RMS results for the 5dB male and female case indicated that the power SS produced less noise, the spectrograms indicate that power SS produces more distortion of the speech as for other values of SNR.

This evidence would tend to suggest that magnitude SS is a better choice for speech enhancement in a non-stationary environment, and this was the approach chosen for the enhancement algorithm that was being developed. This choice was backed up by results in [39], where SS with $\alpha = 1$ was said to sound "*less noisy*".

3.2.2 Choice of subtraction factor β

The second parameter of interest in SS is the subtraction factor β. This determines the percentage of the interfering noise which is to be removed from the corrupted speech. In general the common choice for SS algorithms is $\beta = 1$ as in [6] [94] [76] [40] [66], although recent work [41] [95] [96] [97] [75] has examined a range of β values greater than one to examine their effects on speech enhancement. The justification for this, was that a $\beta \geq 1$ attenuates the noise more than was required, leading to a reduction in the residual noise peaks in the processed speech. It is stated in [95] that for conditions where $\beta > 1$, the residual noise after subtraction would be lower than if $\beta = 1$ had been used.

The disadvantage to this approach is that as the noise is being oversubtracted, the opportunity for corruption of the processed speech is increased. This is heard as audible distortion of the speech. To determine a possible β value, an experiment was carried out where magnitude SS was applied to a male speech/pink noise case with SNR +5dB. Figures 3.2(a) to 3.2(d) show the results of this experimentation, with β ranging from one to four. Higher values of β were not chosen, as it was felt that the maximum useful β value is six.

The spectrograms illustrate the trade off that is inherent with SS between noise reduction and speech corruption. As β moves from one to four in figures 3.2(a) to 3.2(d), it can be seen that while the residual noise did decrease, there was a proportional increase in the distortion of the speech spectrogram which is seen by less of the speech spectrum being present. At the point $\beta = 4$, the majority of the speech data was destroyed. This as a view also presented in [41], where it was reported that at low SNRs altering β resulted in less visible background noise being present in the spectrogram, but there was also a corresponding reduction in the visible

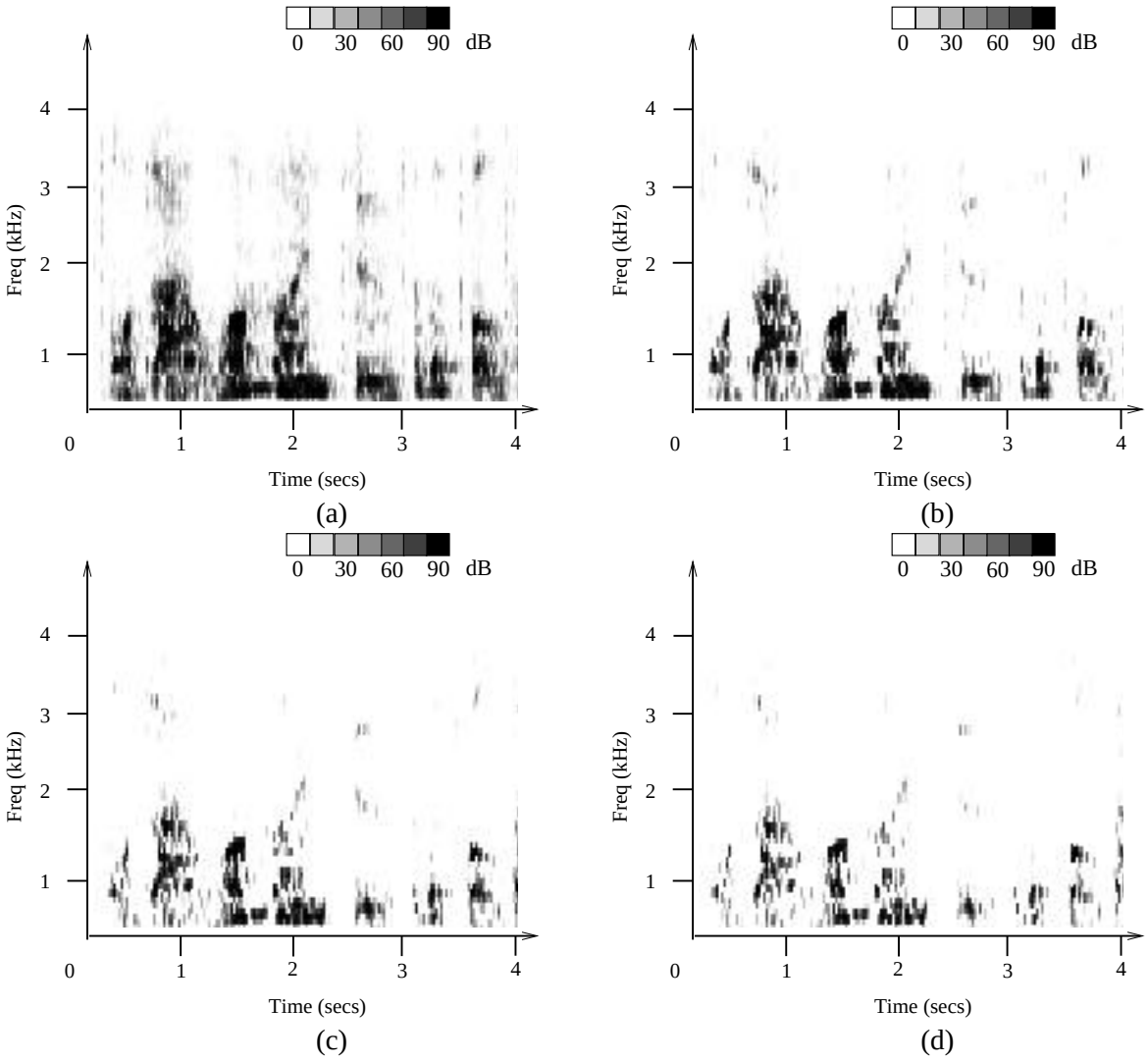


Figure 3.2: *SS with (a)* $\beta = 1$ *(b)* $\beta = 2$ *(c)* $\beta = 3$ *(d)* $\beta = 4$

speech spectrum.

In contrast cases where $\beta < 1$ were not used as all of the possible noise would not be removed from the system, and for these reasons the more popular $\beta = 1$ case was chosen to be implemented in the enhancement algorithm SS.

3.2.3 Dealing with negative values in $|\widehat{S}(m)|$ after SS

In section 1.2, it was explained that during SS it was possible for negative values to occur in the enhanced speech spectrum. As we are dealing with speech magnitude, it is not possible to have negative magnitude values, and these cases must be dealt with. There are three main ways

of doing this.

- Half Wave Rectification.
- Full Wave Rectification.
- Introduction of a Noise Floor.

These techniques also affect the amount of "musical noise" which occurs in the enhanced speech. The topic of musical noise was introduced in Section 2.7, and methods of dealing with it are presented in Chapter 5.

3.2.3.1 Half Wave rectification

In this approach all negative values are set to zero.

$$\widehat{S}(m) = \begin{cases} |Y(m)| - |\widehat{N}(m)| & : \quad |Y(m)| > |\widehat{N}(m)| \\ 0 & : \quad \text{otherwise} \end{cases}$$

This is the simplest form of dealing with negative values, and was used in many approaches such as [6] and [98]. It has the disadvantage of possibly removing speech data where the corrupted value is less than the estimated noise.

3.2.3.2 Full Wave rectification

In this technique the absolute value of the SS replaces all the spectrum points.

$$\widehat{S}(m) = ||Y(m)| - |\widehat{N}(m)|| \tag{3.1}$$

This has the advantage of retaining all of the possible speech information, although the amount of noise reduction is reduced as there is always a residual value, which is difficult to predict. The full wave rectification can lead to points in the spectrum being unnaturally enhanced, causing them to sound out of place, which can corrupt the speech portion of the signal. The residual

background noise can also sound unnatural, as the subtraction and rectification will alter the levels of the noise left in the signal from sample to sample.

3.2.3.3 Noise Floor introduction

In 1979 [99] introduced the idea of replacing negative spectrum values, with a small percentage of the estimated noise. It was reasoned that any low energy speech that could be lost by the subtraction, tends to be unvoiced (noise-like) in nature, and a reasonable estimation of its spectral magnitude could be obtained by replacing the value with a percentage of the corresponding estimated noise sample. This was extended, to define a threshold across the magnitude of the speech below, which all points were replaced (not only negative values) as follows;

$$\widehat{S}(m) = \begin{cases} |Y(m)| - |\widehat{N}(m)| & : \quad |Y(m)| - |\widehat{N}(m)| > \eta\widehat{N}(m) \\ \eta\widehat{N}(m) & : \quad \text{otherwise} \end{cases}$$

where η denotes the percentile which determines the height of the noise floor. The choice of η has a great bearing on the resultant spectrum of the enhanced speech. It not only determines the value of the point replacing any which fall below the threshold, but also determines how many points are subject to the spectral subtraction enhancement, and the amount of residual noise which is added to the system. Obviously, the introduction of such a noise floor can be counterproductive, as it destroys the work of the spectral subtraction. For this reason, only very small values of η were examined to prevent this. In [99] a value of 0.02 was suggested, and [95] examined this along with 0.001 and 0.008. Each of these were claimed to give their own advantages for the enhancement of speech, although the experimentation in these papers was only performed on stationary noise conditions. To examine if these η values would be of use in a non-stationary noise condition, a SS experiment was performed ($\alpha = 1, \beta = 1$), using these and a couple of intermediary values (0.003 and 0.015), along with the half and full wave rectification cases. From this, it could be determined which was the best method of dealing with negative values in non-stationary noise conditions. The signals operated on were a female/pink noise at -5dB and a male/music at 0dB, and the same section of speech was examined as previously. A selection of spectrograms from this experimentation are presented here, the full set can be found in Appendix A as figures A.1 to A.4

From figures 3.3(a) and 3.3(b), the effects of full wave as opposed to half wave rectification can

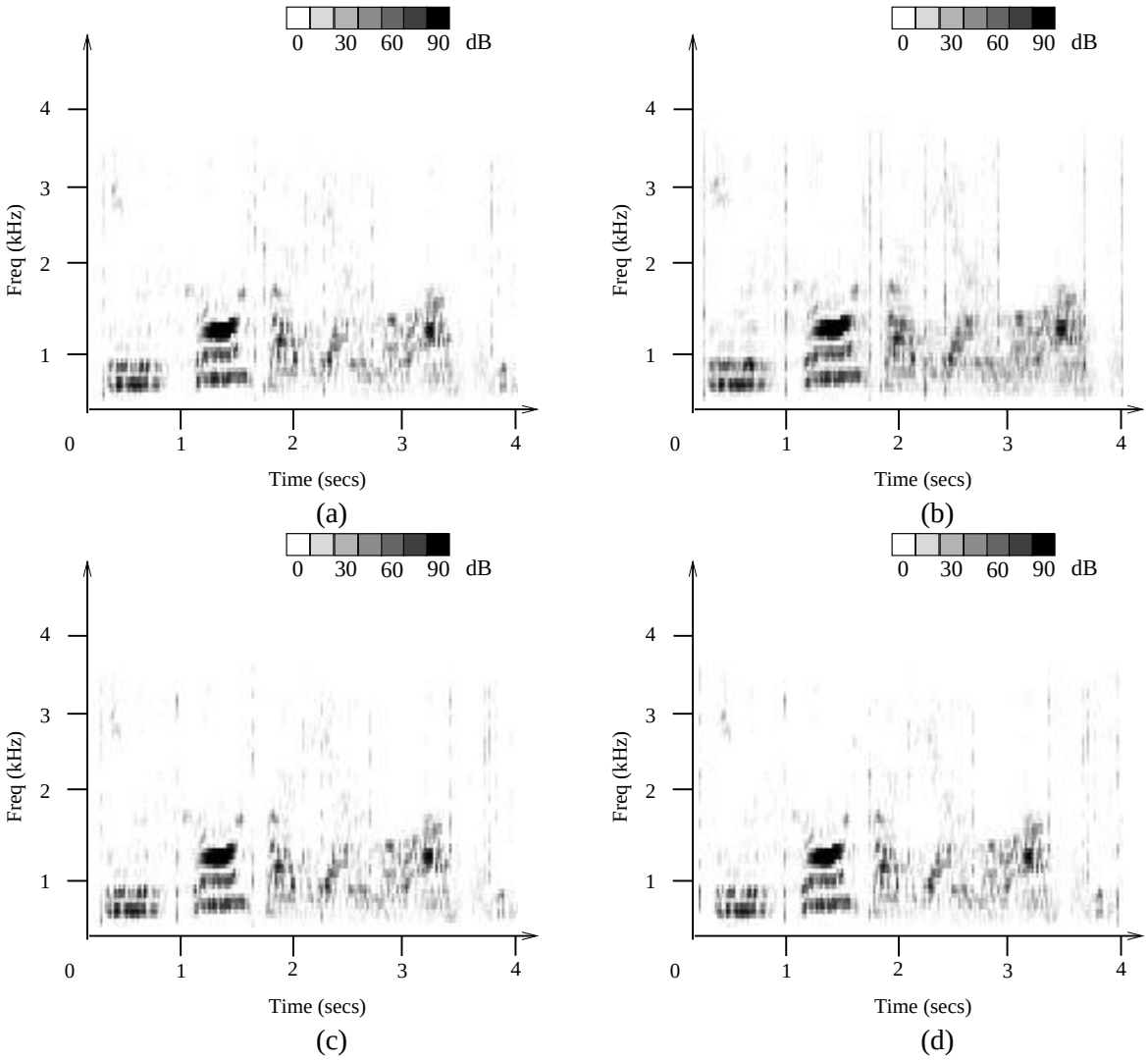


Figure 3.3: *Female Voice with (a) Half Wave rectification (b) Full Wave rectification (c)* $\eta = 0.001$ *(d)* $\eta = 0.02$

be clearly seen. While less of the speech information is lost, there is an increase in the amount of non-speech information evident, even in this small portion of the speech. The same is also true of figures 3.4(a) and 3.4(b). Comparing these results to figures 3.3(d) and 3.4(d), which represent the largest η value studied, it can be seen that full wave rectification still introduces more non-speech elements, and does not seem to be the best method to choose.

The choice between η values is more complex, as there is a subtle trade off between increased background noise, and speech corruption. Table 3.3 shows the average RMS magnitude errors which occur from the experiments conducted. From these statistical results, it could only be surmised that the more η is increased, then the closer to the original speech the replacement

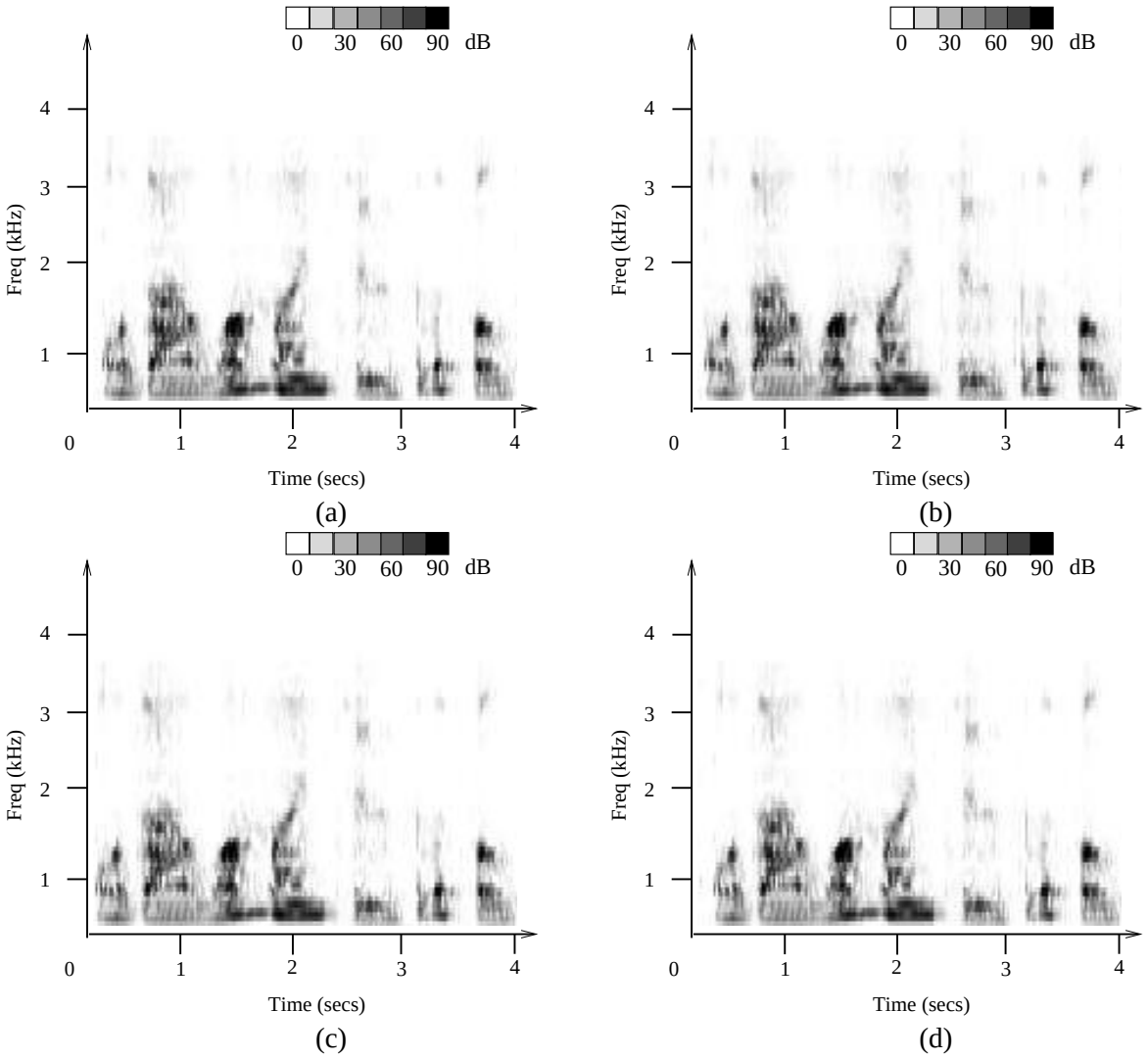


Figure 3.4: *Male Voice with (a) Half Wave rectification (b) Full Wave rectification (c)* $\eta = 0.001$ *(d)* $\eta = 0.02$

value becomes. This trend reverses when the η value is so high, that a high percentage of the background noise is replaced, defeating the purpose of the spectral subtraction. To illustrate this, cases of $\eta = 0.1, 0.2$ and 0.5 were also examined. It can be seen from the RMS results, that once η raises above 0.1, the distortion increases considerably. The only way to choose between the other values was by informal listening tests. In general the value $\eta = 0.008$ perceptually presented the best trade off between the noise and speech distortion. Going higher than this led to an increased perception of the background noise, and less than this caused more severe speech distortion.

This completed the work required for the basic spectral subtraction routine, which could be

	Male Speech	Female Speech
Operation	RMS percentage error	RMS percentage error
Half Wave	1.611602	1.299322
Full Wave	1.473209	1.197228
$\eta = 0.001$	1.610788	1.298232
$\eta = 0.003$	1.609176	1.296067
$\eta = 0.008$	1.605259	1.290746
$\eta = 0.015$	1.599991	1.283596
$\eta = 0.02$	1.596434	1.278748
$\eta = 0.1$	1.570844	1.240890
$\eta = 0.2$	1.635972	1.322418
$\eta = 0.5$	2.376257	2.211218

Table 3.3: *Average RMS magnitude percentage error results for changing noise floor percentiles.*

described by;

$$\widehat{S}(m) = \begin{cases} |Y(m)| - |\widehat{N}(m)| & : \quad |Y(m)| - |\widehat{N}(m)| > 0.08\widehat{N}(m) \\ 0.08\widehat{N}(m) & : \quad \text{otherwise} \end{cases}$$

3.3 Integration of the perceptual criterion

The final aspect of the SS, was the integration of the masking threshold information into the decision making process for the subtraction. It was discussed in Section 2.6 why masking thresholds are important elements to integrate into SS work. The ability to reduce noise to a perceptually unimportant level brings an extra measure of accuracy to the speech enhancement algorithm.

In order for this, the perceptual element of the algorithm only looks at the speech after the subtraction process has taken place. From this, it can be determined which points are audible and of interest, and which are inaudible. The perceptual work integrated into the SS was based on that of Johnston [59]. Figure 3.5 shows the steps inherent in the calculation of the auditory masking thresholds.

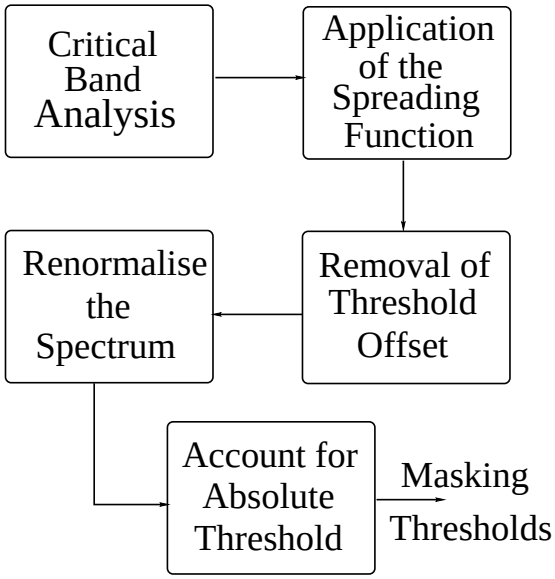


Figure 3.5: *Flow Diagram of the perceptual process.*

3.3.1 Critical band analysis

The first section of the perceptual analysis was to represent the auditory signal in a way similar to the human ear. In section 2.6.3 the critical band was defined to be separation that must exist between two frequencies, in order for them to be heard individually. Given the short time DFT magnitude of the enhanced speech $|\widehat{S}(m)|$, the signal was transformed into the power spectrum.

$$P(m) = |\widehat{S}(m)|^2 \tag{3.2}$$

The 128 samples of the power spectrum were then partitioned in critical bands (CB), according to frequency as shown in table 3.4, for the $0 - 4$kHz signals operated on in this thesis.

The range of pitches as perceived by the human auditory systems, is divided up into intervals corresponding to CB or barks. The bark scale was defined by Zwicker [100], and can be calculated for any audible frequency using the equation:

$$\mathrm{b} = 13 \times \arctan\left(\frac{0.76 \times \mathrm{fre}}{1000}\right) + 3.5 \times \arctan\left(\frac{\mathrm{fre}}{7500}\right)^2 \tag{3.3}$$

where b, fre are the number of barks, and the frequency in hertz respectively. The CB bound-

Critical Band Number	Lower Limit (Hz)	Center Frequency (Hz)	Upper Limit (Hz)
1	0	50	100
2	100	150	200
3	200	250	300
4	300	350	400
5	400	450	510
6	510	570	630
7	630	700	770
8	770	840	920
9	920	1000	1080
10	1080	1170	1270
11	1270	1370	1480
12	1480	1600	1720
13	1720	1850	2000
14	2000	2150	2320
15	2320	2500	2700
16	2700	2900	3150
17	3150	3400	3700
18	3700	4000	4400

Table 3.4: *Critical Band Centers and Limits.*

aries and the bark scale are related. The integer values of the bark scale, are taken to be the critical band boundaries, and another name for the bark scale is critical band rate.

In general CB analysis of the signal can be performed by passing the signal thorough a group of Gaussian shaped band pass filters, which are specified by their center frequency, and bandwidth as in [68].

$$B_k(j) = \exp\left[-1.5\left(\frac{j - F_k}{BW}\right)^2\right] \; for\; k = 1, 2...18 \tag{3.4}$$

where $B_k(j)$, k, F_k, BW, $j \in (a_k, b_k)$ are the frequency response of the CB, the CB number, the center frequency, the bandwidth of the CB filter, and the frequency sample of interest which lies between the upper or lower limit of the frequency range of the band pass filter, respectively.

In [59] the CB filtering is simulated by summing up the energy in each CB giving;

$$E_k(m) = \sum_{k=a_k}^{b_k} P(m) \tag{3.5}$$

where $E_k(m)$, a_k and b_k are the energy, lower and upper limits of CB k.

This then produces the power spectrum in the Bark domain which looks like figure 3.6. This shows a window of male speech in the Bark domain.

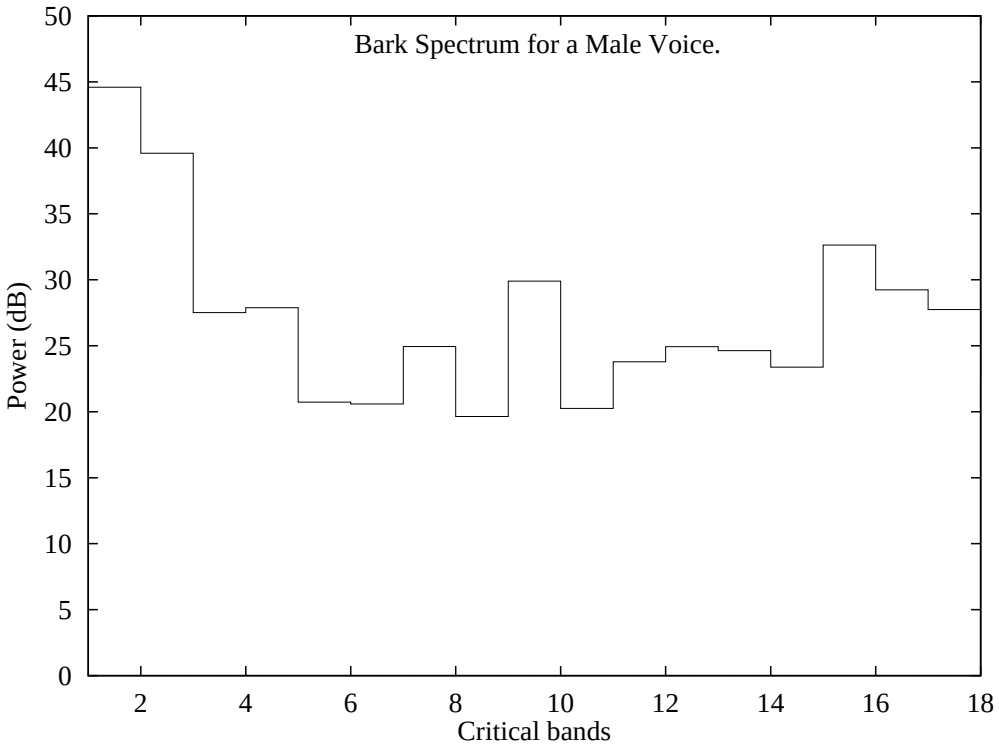


Figure 3.6: *A typical bark spectrum for male speech.*

In this figure, it can be seen that the bark domain representation is constant over each CB. This is due to the energy summation which the CB analysis uses.

3.3.2 Application of the Spreading Function

Once the conversion to the bark domain has been achieved, the next step is to apply the spreading function. This determines how much an excitation in one CB will affect other CBs. With a mobile phone, there is the potential to experience interference right across the telephone bandwidth and so masking can occur across all 18 CBs. This means the spreading function has to be calculated for all conditions that satisfy;

$$\mathrm{abs}(l-k) \leq 18 \tag{3.6}$$

where k and l are the bark frequency of the masked signal and the masking signal respectively.

The spreading function is essentially a triangular shaped function as shown in figure 3.7. The only choice in the function centers around the fall off rate of the spreading slopes. In figure 3.7 these can be seen as $+25$dB and -15dB per bark. The spreading function used comes from Schroeder [58] which had slopes of $+25$dB and -10dB.

$$10\log_{10} Sp(k) = 15.81 + 7.5(k + 0.74) - 17.5(1 + (k + 0.474)^2)^{\frac{1}{2}}\text{dB} \tag{3.7}$$

where k = CB number.

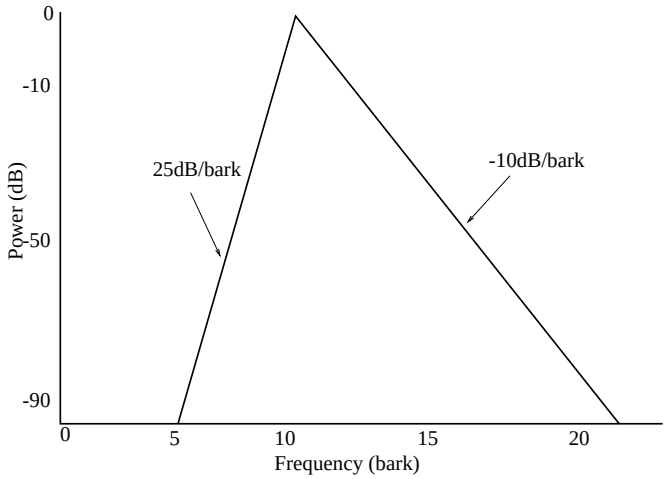


Figure 3.7: *A typical spreading function*

The coefficients of the spreading function were held in an 18 by 18 matrix, which determined all the possible interactions between the masking signal and masker.

$$\begin{pmatrix} Sp_{1,1} & Sp_{1,2} & \ldots & Sp_{1,18} \\ Sp_{2,1} & . & . & . \\ . & . & . & . \\ Sp_{18,1} & \ldots & \ldots & Sp_{18,18} \end{pmatrix}$$

The convolution of this spreading function matrix and the CB excitation, was performed in the frequency domain by matrix multiplication with the result of (3.5), to get the spread CB

spectrum C_k.

$$C_k = Sp_{jk}B_k \tag{3.8}$$

An example of the spread bark spectrum can be seen in figure 3.8, which is visibly different to figure 3.6 due to the application of the spreading function.

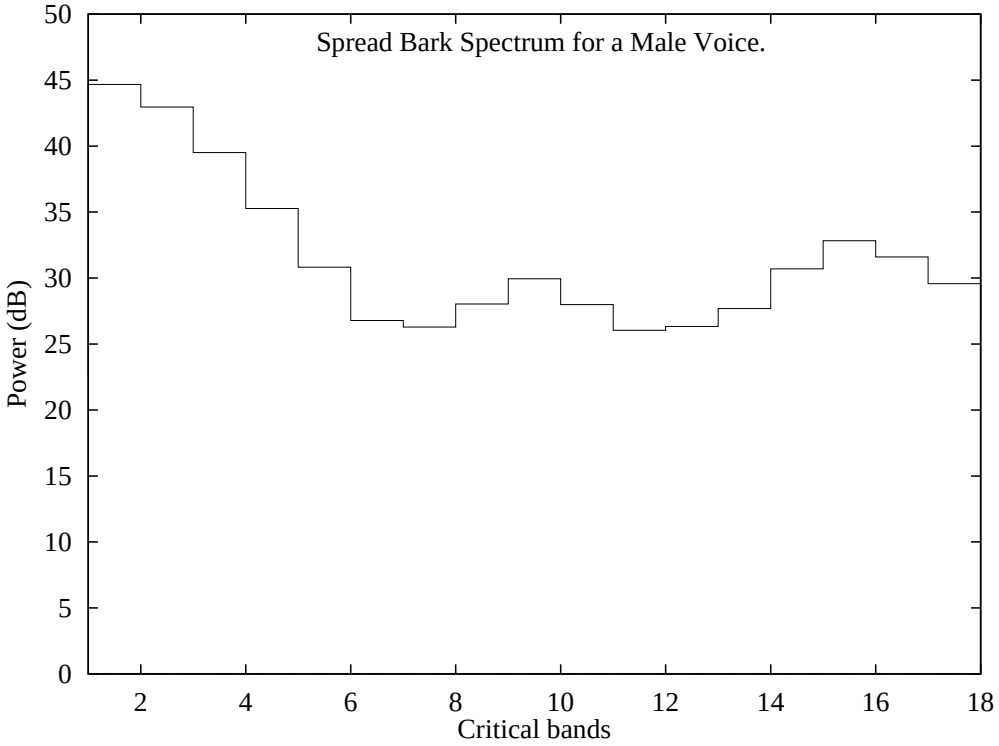


Figure 3.8: *A typical spread bark spectrum for male speech.*

3.3.3 Calculation of Thresholds and Removal of Offset

Having established in section 2.6.3 that the cochlea was the only part of the auditory system which had any affect on noise reduction, the next step was to determine the masking thresholds related to the excitation the cochlea receives. There are two types of masking threshold, Noise Masking Tone (NMT), and Tone Masking Noise (TMN). Each criterion present a different threshold level. For a TMN it is stated in [59] this was represented as;

$$T_k = C_k - (14.5 - k)\ \mathrm{dB} \tag{3.9}$$

where T_k is the estimated masking threshold.

and for NMT;

$$T_k = C_k - 5.5 \text{ dB} \tag{3.10}$$

When the spread bark spectrum was presented, a decision had to be made whether the signal was tone-like or noise like. This determined whether the threshold for the CB required the application of equation (3.9) or (3.10). In order to do this the, spectral flatness measure (SFM) was used to determine the *tone-likeness* of the signal. This was calculated per CB by;

$$SFM_k \text{ dB} = 10 \log_{10} \left(\frac{Gm(k)}{Am(k)} \right) \tag{3.11}$$

where $Gm(k)$ and $Am(k)$ are the geometric mean and arithmetic mean of the spectrum over CB k.

The SFM_k dB from (3.11) was then used to derive the tonality constant γ;

$$\gamma = \min \left(\frac{SFM_k \; dB}{SFM_{\text{dBmax}}}, 1 \right) \tag{3.12}$$

where $SFM_{\text{dBmax}} = -60$ dB is the reference level for a tone like signal.

The comparison of SFM_k dB to SFM_{dBmax} gave γ a value which ranged between $0 - 1$. If $SFM_k \text{ dB} = 0$ then $\gamma = 0$ meaning the signal was noise-like, a $SFM_k \text{ dB} = -60$ would mean the signal was tone-like. The tonality parameter γ was used to weight between equations (3.9) and (3.10) to provide an offset which was subtracted from the estimated threshold.

$$Off_k = \gamma(14.5 + k) + (1 - \gamma k) \tag{3.13}$$

This step was performed to allow for a prediction of the TMN and NMT thresholds at all frequencies in the spectrum. As an example [59], stated that speech signals generally had a $SFM = -20 \rightarrow -30$dB which related to a $\gamma = 0.3 \rightarrow 0.5$. The spread masking thresholds were then calculated by removing this offset from the spread bark spectrum.

$$Ts_k = 10^{\log_{10}(C_k - \frac{Off_k}{10})} \tag{3.14}$$

Ts_k is the spread masking threshold for CB k.

Frequency (Hz)	MAP (dB SPL)
100	33
150	24
200	19
300	13.3
400	9.6
500	7.7
700	7.5
1000	6.8
1500	8.0
2000	13.0
2500	15.6
3000	12.5
3500	10.9
4000	10.3

Table 3.5: *Minimal audible pressure values.*

3.3.4 Renormalisation

During the convolution operation of equation (3.8), the bark spectrum was multiplied by a non-trivial gain function which was due to the interaction of the masking threshold across the 18 CBs. This had to be undone, to bring the thresholds back into the bark domain, but was too complex a function to simply invert. It is common practice to take a much simpler route, and model this gain by a renormalisation, using the DC gain in each CB as the normalisation factor, as is shown in equation (3.15)

$$Tb_k = \frac{Ts_k}{DCgain} \tag{3.15}$$

3.3.5 Comparison to Absolute Threshold

The final point of the threshold calculation, was comparing the calculated thresholds to the absolute thresholds for the human ear. These absolute thresholds determine the minimum signal which is audible to the human ear, and when compared to the calculated values, ensure that no threshold was set too low. Much of the work on this was presented in [101], where an examination was made of the minimal perceptible changes in pitch, and its effect on masking threshold levels. For the perceptual system which was being integrated into the SS the absolute thresholds came from [102] as shown in Table 3.5.

This completed the threshold calculation for the SS algorithm. As shown in figure 3.9, these thresholds were compared to the estimated clean speech. If the spectrum value was above the threshold it was unaffected, if below the threshold, the noise floor approach of section 3.2.3.3 was applied, and a small percentile of the reference noise replaced the non-audible quantity. This caused the speech enhancement to only operate on those points which were audible.

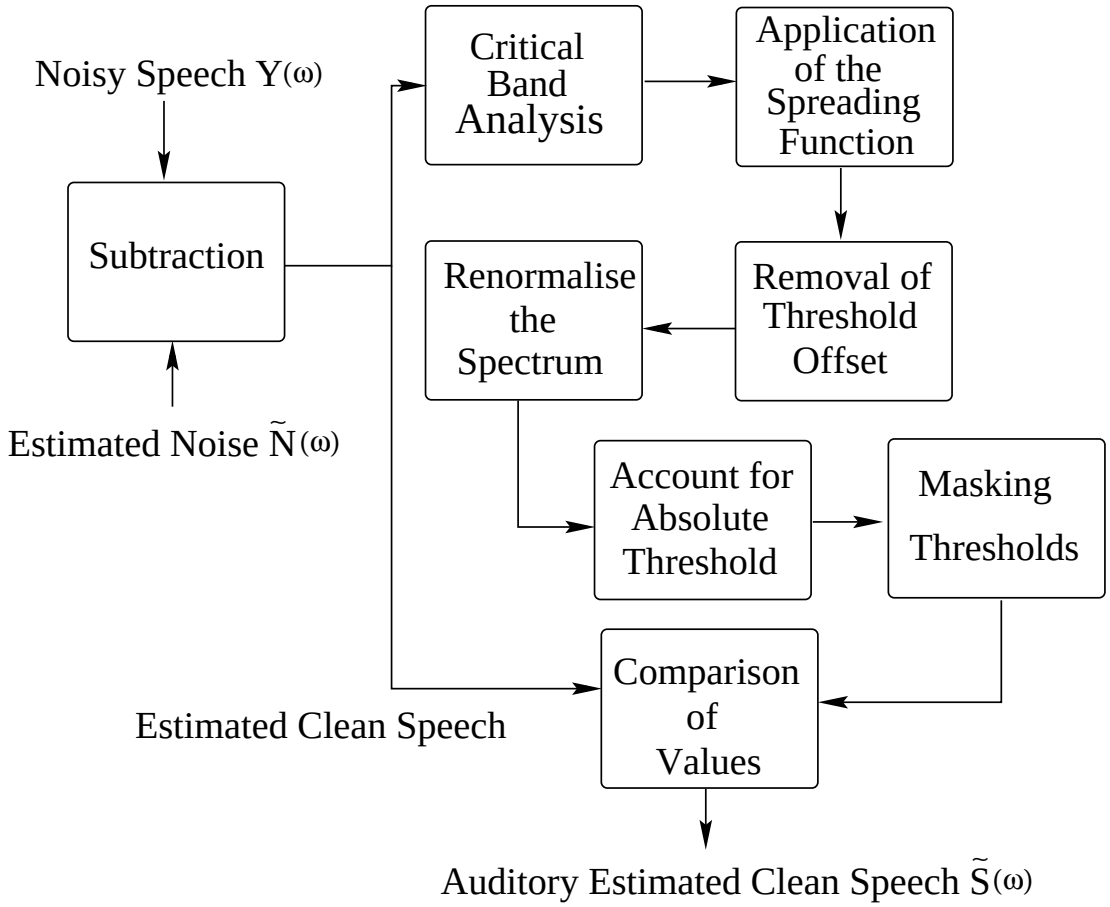


Figure 3.9: *Flow Diagram of how the perceptual criterion are integrated into SS.*

3.4 Summary

In this Chapter the auditory enhanced SS algorithm, which was the basis of the speech enhancement algorithm was examined. The choice of DFT window, its size and overlap were described, and it was seen that the choice of α and β play an important aspect in the residual noise and distortion of the speech which occurred. It was discovered that for low SNR cases, magnitude

SS was more suitable than power SS, and that oversubtraction of the noise was not advantageous, leading to the choice of $\alpha = 1$ and $\beta = 1$. The advantages and disadvantages of half wave rectification, full wave rectification, and noise floors were discussed, and through testing it was shown that the noise floor approach produced speech with fewer artefacts, which led to its choice. It was shown that while SS was a useful algorithm, no account was made of how the human ear perceived the signal, and this led to the implementation of masking thresholds. The theory behind the calculation of these thresholds was provided, and led to a SS algorithm onto which other portions could be added, to deal with the particular case of non-stationary noise sources for a mobile phone user. The SS technique relies on the accuracy of the noise estimate to enable it to produce speech with low residual noise and natural sounding input, the next Chapter investigates how such an estimate may be produced.

Chapter 4
Obtaining an Accurate Noise Estimate.

In this Chapter, the importance of an accurate noise estimate is examined and the effects of inaccurately modelling the background noise in SS are described. Two new algorithms to improve the estimation for non-stationary noise are presented, and subjective listening tests determine the performance of these compared to some other techniques. The work presented in this Chapter is an extension of that published in [103]

4.1 Criteria for a Noise estimation algorithm.

In section 2.7, the importance of an accurate noise estimate was outlined. In figure 3.9, it was seen that the estimate of the noise, $|\widehat{N}(m)|$, was a vital part of the SS process, and (as stated in section 2.7) determined the:

- amount of noise reduction (increase in SNR);
- influence of musical noise (corruption of the resultant speech making it less intelligible);
- loss of speech data (loss in quality of speech).

As the application of the work presented is enhancing mobile phone speech, corrupted by non-stationary noise, there are a number of criterion which define the outline of the noise estimation and they are:

- low computational complexity, so that the user does not have a long time lag until the enhanced speech is presented to the loudspeaker;
- the ability to perform in a variety of noise environments;
- the reduction of erroneously selecting speech points as noise and vice-versa.

If one of these criterion is not met, then the speech out of the enhancement algorithm will suffer from a combination of the corruption outlined above. In order to define the parameters of the noise estimation technique, some assumptions were made:

1. the noise was additive, non-stationary and uncorrelated from sample to sample;
2. the speech and noise were uncorrelated;
3. there was a reference microphone, which monitors the environmental noise, but had no speech cross talk present.

From assumptions 1 and 2, it could be seen that any technique which could satisfy these requirements would also be able to remove stationary noise, as this was a simpler noise case to deal with. Point 3 states the requirement for a possible two sensor system, as opposed to only using the microphone on the mobile phone handset.

4.1.1 One or Two Sensor System?

In Section 2.7, there was an outline of some of the past work performed on the topic of noise estimation, as part of speech enhancement. It could be seen, that a great deal of the spectral approaches fell into the category of either one or two sensor designs. The choice between these impacted greatly on the accuracy of the noise estimation, which can be provided for the enhancement.

Single sensor systems rely on either explicit [72], or soft decision [98] speech/pause detectors. The main disadvantage of using this, especially with single sensor approaches, is the interval between updates. During this time, the noise estimate is held at the level of the last update, and it may be a number of seconds until the next update. For stationary noise conditions, this is satisfactory, as during the interval there will be no alteration of the noise conditions. Application of the same theory to non-stationary noise is not as suitable, as the background noise may have changed drastically before the next speech pause. Keeping the noise estimate the same during this interval would lead to a highly inaccurate estimate being applied to the SS process, and result in a degradation of the enhanced speech.

For this reason, an approach using two sensors was chosen, and this meant that at all times the noise estimation has an input from a microphone, which could provide as close a signal to the

actual noise as possible. The positioning of the microphones on a typical mobile phone handset is shown in figure 4.1.

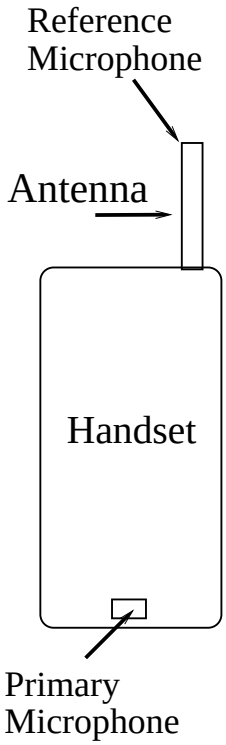


Figure 4.1: *The proposed positioning of the reference and primary microphones.*

The reasoning for this is similar to the reference microphone justification of [9] and [11], in that the job of the reference microphone is to capture the general noise environment. Here though, the element of non-stationarity rises, as it can be possible that the noise presented to the primary microphone may be different to that at the reference. The use of two microphones always allows at least one to be taking an estimate of the background noise, although if possible the primary can give an estimate that is closest to the corruption any speech experiences. This was the basis for the noise estimation technique which was developed.

4.2 The NEAH technique

The introduction of a noise estimation technique expanded the speech enhancement algorithm to the point shown in figure 4.2

Early on in the development of the speech enhancement algorithm, it was felt that a new approach to noise estimation should be taken. This was especially true, as most of the work in

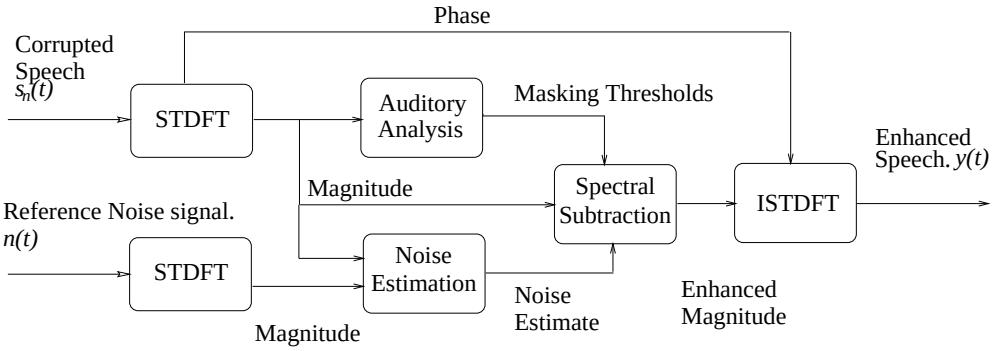


Figure 4.2: *The enhancement algorithm including auditory enhanced SS and noise estimation.*

this area had centred on a stationary noise condition. Of primary concern for the non-stationary case, was the incorrect detection of noise as speech and vice-versa. As speech/pause detection plays a significant role, allowing the system to produce an accurate estimate, a low complexity solution which can produce a good estimate was required. With the two microphone approach shown in figure 4.1 it is possible to combine a simple speech/pause detector with a secondary microphone allowing a background noise estimate even during speech. The estimation could also be improved, if the technique could verify whether a section of signal was speech or noise before applying the estimation. This was the theory behind the Noise Estimation with A Hangover (NEAH) technique which is depicted in figure 4.3

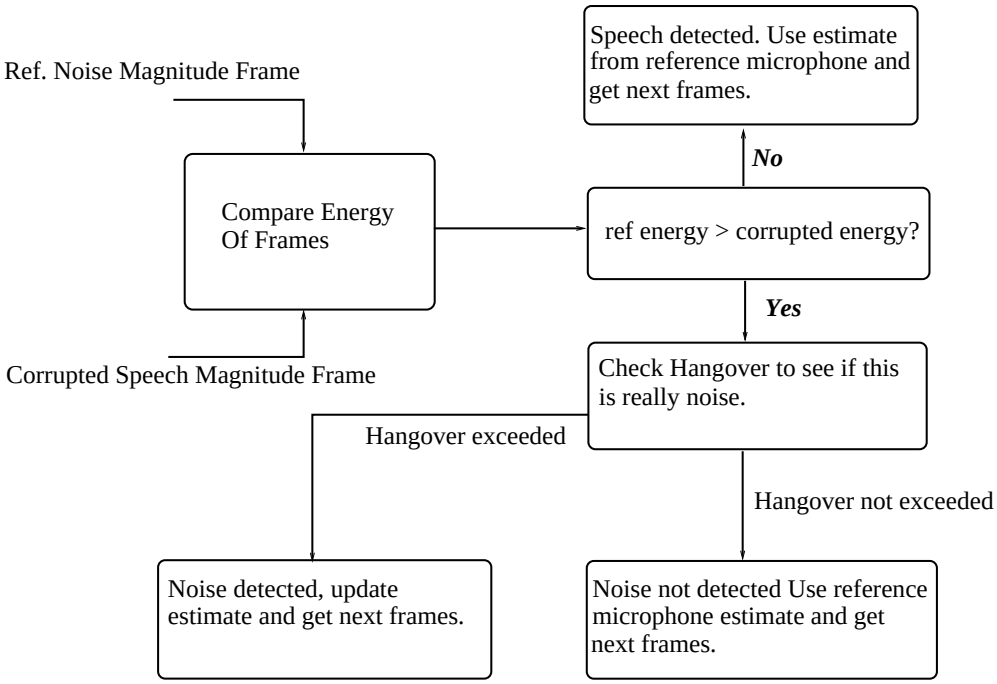


Figure 4.3: *Flow Diagram of the NEAH technique.*

4.2.1 Energy Estimation

The first step was to implement a simple speech/pause detector. While there were many kinds of detection which could have been chosen [104][35][36], a simple energy detector was used, as the interaction between the speech and background noise will result in an energy mismatch between the reference and primary microphone.

For each frame of speech the energy of the primary and reference channels were calculated:

$$E_{prim} = \sum_{n=0}^{127} |Y_n(m)|^2 \tag{4.1}$$

$$E_{ref} = \sum_{n=0}^{127} |N_n(m)|^2 \tag{4.2}$$

where E_{prim}, E_{ref}, n are the energy in the primary frame, reference frame, and sample number in frame respectively.

The speech/pause energy detector was classified by:

if $E_{prim}(m) > E_{ref}(m)$ speech and noise present.
if $E_{prim}(m) \leq E_{ref}(m)$ possibly only noise present.

4.2.2 Hangover mechanism

If the second condition of possibly only noise being present was entered, then the technique waited until the next frame of the signal was analysed, to determine whether this was speech or noise.

if $E_{prim}^{next}(m) > E_{ref}^{next}(m)$ then it is not noise so use last update.
if $E_{prim}^{next}(m) \leq E_{ref}^{next}(m)$ then it is definitely noise and update.

This delay or *hangover* was kept to only one frame, in order to ensure that the technique could track the non-stationary noise. A one frame hangover also just touched the end of the time

limit over which the speech could be considered to be stationary. It was also noted that upon listening to the NEAH technique with longer hangovers, the accuracy of the noise estimate reduced, leading to greater corruption of the speech produced by the enhancement algorithm.

Having determined a simple detection system which could cope with the non-stationarity, the next aspect to examine was the way in which to update the estimate of the noise. For normal speech/pause detection the noise update is computed as:

$$\widehat{N}^{New}(m) = (1-p)\widehat{N}^{last}(m) + Y(m) \tag{4.3}$$

where $\widehat{N}^{last}(m)$, $\widehat{N}^{New}(m)$, $Y(m)$ and p are the previous estimate, the new estimate, the primary frame and the forgetting factor respectively. Generally the forgetting factor is set between $0.5 - 0.9$.

Equation 4.3 uses the forgetting factor p to keep a track of the history of the previous noise estimates. This weights the update, so that different portions of the present noise and past update are used to compute the present output. The more non-stationary the environment then the less emphasis is placed on the past noise estimates.

In this version of the NEAH technique, the speech/pause detection was used to determine which microphone the noise estimate should be taken from. The justification for this being that when there was definitely noise present in the primary microphone, this should be the estimate of the noise, as this is closest to the interference which the speech shall encounter. At other times the signal from the reference microphone was used to take a new estimate.

To compare this with the approach given typified by equation (4.3), the NEAH technique was constructed in two variants, one with the update method outlined above, and one using equation (4.3) for the update. In this, the signals which were to be updated came from the reference microphone, the primary being used only for the detection of speech/pause information. Two test signals were used, a male or female voice, each corrupted with pink noise with SNRs ranging from $-10 \rightarrow 5$dB. Figures 4.4(a) and (b) show the change in SNR resulting from the application of each of these techniques. In these figures the plot *NEAH1* represents the technique which chooses between the two microphones, *NEAH2* represents the estimation brought about by equation (4.3)

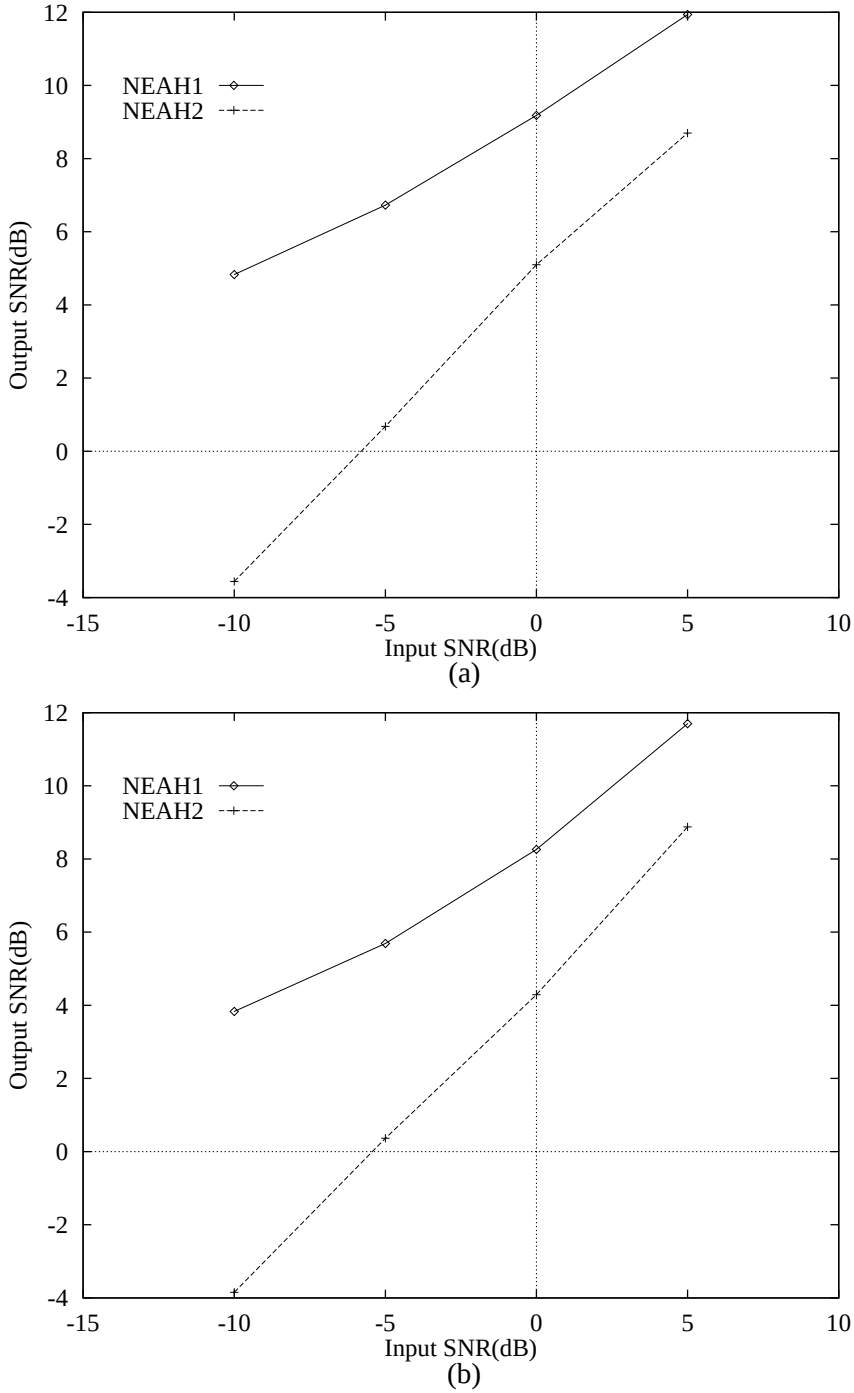


Figure 4.4: *NEAH SNR improvement graph for (a) Female Voice (b) Male Voice*

It can be seen from both figure 4.4(a) and (b) that the SNR of the output speech increases as the SNR of the input speech increases. This was of course expected as there was less noise being introduced into the algorithm. It was also apparent from figure 4.4, that the NEAH technique which chooses between the two microphones, exhibits better SNR improvements than the estimation using the standard method of equation (4.3).

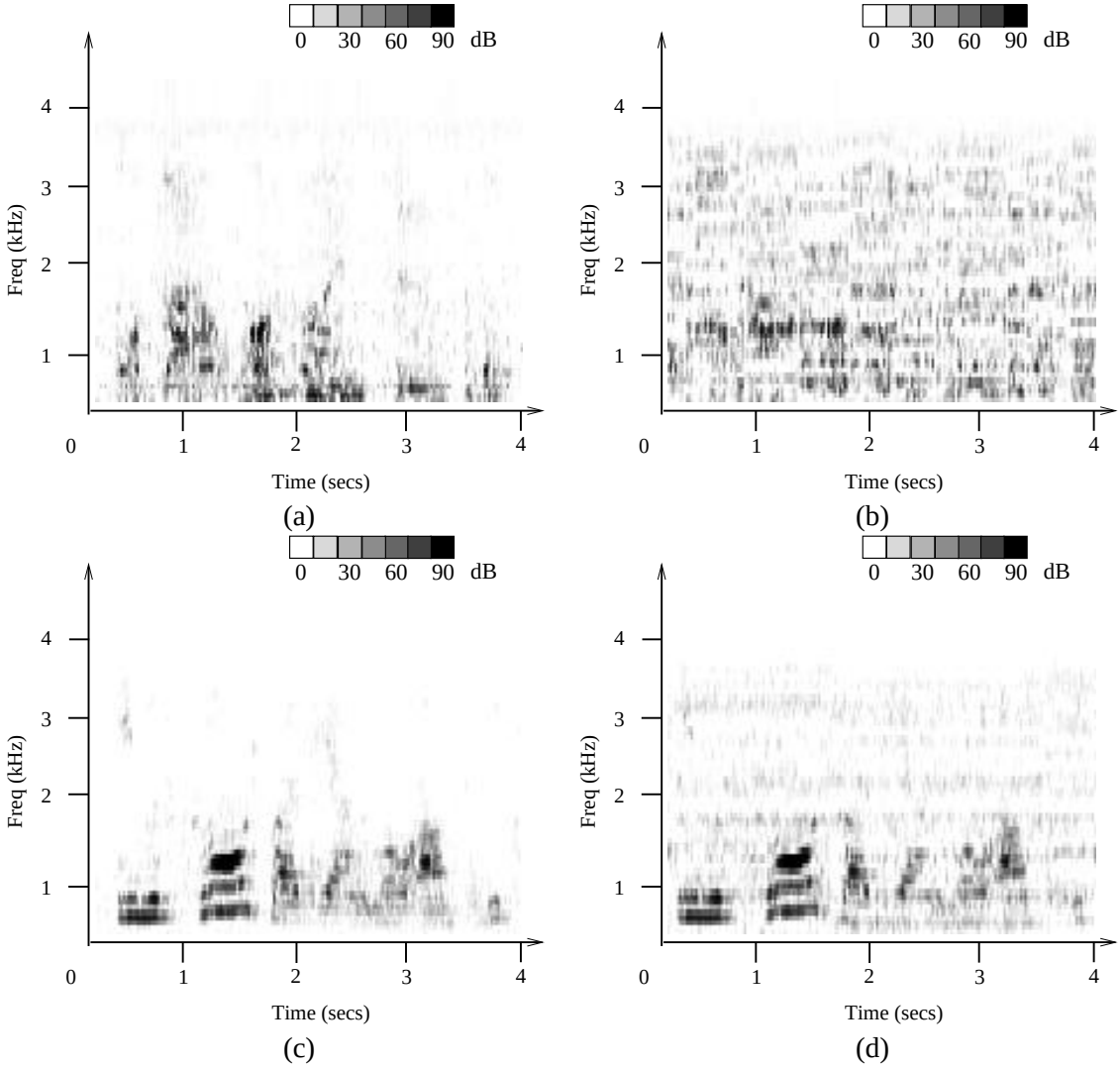


Figure 4.5: *Spectrograms for (a) Two microphone switching* NEAH1 *Male (b) Update equation* NEAH2 *Male (c) Two microphone switching* NEAH1 *Female (b) Update equation* NEAH2 *Female*

An interesting point, is that this is one of the few cases when SNR improvement can actually give quantification of the quality of the speech which an technique may output. This occurs because the technique being studied was the actual estimation of the background noise, and which approach was better. Obviously, a bad background estimate will result in poor speech

enhancement in terms of intelligibility and listenability.

To justify this, figures 4.5(a-d) are the spectrograms of speech which results from experimentation on signals with the conditions: Male Voice/Pink Noise at -10dB (figure 4.5(a,b)) and Female Voice/Pink Noise at 5dB (figure 4.5(c,d)).

These spectrograms show results from experiments using the same portions of the male and female speech that were presented in Chapter 3. It can be seen that the amount of noise which remains after the microphone switching case, is considerably less than for the update equation case. This can explained as the non-stationary noise was uncorrelated from sample to sample, meaning there was no direct relationship between the current, and previous noise estimates. Any attempt to use such a relationship would only result in inaccuracy in the noise estimate, which can clearly be seen in figure 4.5(b) and (d).

SNR and MOS listening test results for cases of speech corrupted with pink noise, a second speaker, a 1kHz tone, and music are presented in section 4.4 to provide a fuller comparison of the respective performance of this noise estimation technique.

4.3 The PANE Noise estimation technique

During the development of the NEAH technique, it became apparent that while noise estimation was an important aspect of SS, little research had been conducted into improving it. What was very striking was that, in all the speech enhancement algorithms developed, only one noise estimation technique was used in each case. This had the disadvantage, that while a particular estimation technique may work well for a particular noise scenario, it may not work well in another. As the purpose of this work was to develop an technique which would perform well, no matter the environment the mobile phone was used in, it became apparent that the NEAH technique on its own may not be an best solution.

From this, it was decided to examine the benefits of estimating the noise through a variety of techniques in parallel. This would have the advantage of minimising the errors which could occur from a single technique, and potentially exploit the performance of each estimation technique in its favoured environment.

It was important to choose estimation techniques which had similar computational complexities. As the noise estimation technique was running these in parallel, the performance would

be limited by the slowest method chosen, which for a mobile phone must not be long as previously stated. It must be noted though, that at this point the focus was on developing a technique which could be further refined for use in the mobile phone case, rather than aiming for all out performance at this stage. Along with the NEAH technique, four other estimation techniques were chosen to integrate into this technique. They were chosen not only for their computational complexity, but the manner in which the estimate was obtained. The structure of this Parallel Noise Estimation technique (PANE) is shown in figure 4.6.

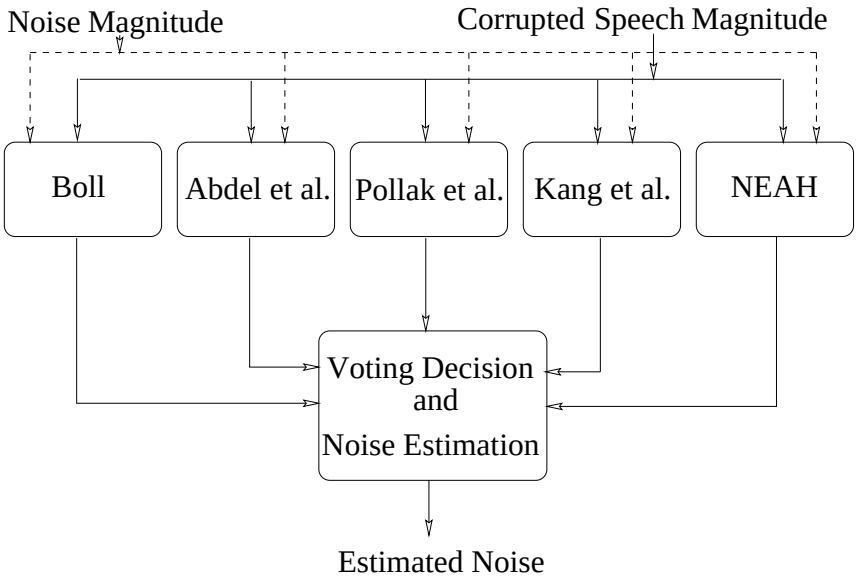


Figure 4.6: *Flow Diagram of the PANE technique.*

The PANE technique sent the corrupted speech and reference noise signals to the five techniques in parallel. When the majority of the techniques decided that there was a frame of noise, an update was performed. The specifics of the decision making process, and the update mechanism are described after an explanation of the constituent estimation techniques. As well as an explanation of the noise estimation, the speech enhancement algorithms they were implemented in are described and their performances will be compared to the NEAH and PANE techniques.

4.3.1 Boll's Estimation Algorithm

In [6] Boll proposed one of the best known noise suppression techniques for speech, which used SS of the form:

$$\widehat{S}(m) = \left[|Y(m)| - |\widehat{N}(m)| \right] \tag{4.4}$$

where $\widehat{S}(m)$, $Y(m)$, and $\widehat{N}(m)$ are the estimated clean speech, the corrupted speech, and the estimated background noise respectively.

For the estimation of the noise $\widehat{N}(m)$ Boll proposed the use of the average value for each frequency bin:

$$\mu(m) = E\left[|N(m)|\right] \tag{4.5}$$

where $E\left[|N(m)|\right]$ is the expected value of the background noise.

The averaging condition took place during non-speech activity, which was determined by energy comparison giving SS:

$$\widehat{S}(m) = \left[|Y(m)| - \mu(m)\right] \tag{4.6}$$

and was followed by half wave rectification.

Boll realised that the more non-stationary the noise spectrum became, then the less useful this technique would be.

> "If the noise magnitude spectrum is changing faster than can be computed, them time averaging to estimate $\mu(m)$ cannot be used. Likewise, if the expected value of the noise spectrum changes after an estimate of it has been computed, then noise reduction through bias removal will be less effective or even harmful i.e., removing speech where little noise is present"

It is interesting to note that the SS used after the noise estimation was magnitude based, a choice which had already been made in section 3.2.1.

4.3.2 Abdel, Mokhtar and Ezz-Al-Arab

In [95] a method to enhance speech in vehicles, improving the efficiency of mobile phone encoders, was presented. The premise was, that the encoders for speech transmission in mobile phones require the speech signal to be noise free to work well. Again this was a STDFT based

method but unlike those described previously this used power spectral subtraction:

$$|\widehat{S}(m)|^2 = \left[|Y(m)|^2 - est|N(m)|^2\right] \tag{4.7}$$

The value used for the background noise estimate $est|N(m)|^2$, in this technique came from direct measurements of the environmental noise, while there was non-speech activity. Once the measurement had been made the background noise was oversubtracted by a value of three, and it was suggested that a small noise floor be implemented, to reduce the effects of musical noise artefacts.

$$P_{\widehat{S}}(m) = \begin{cases} P_Y(m) - 3P_{\widehat{N}}(m) & : \quad P_Y(m) \geq 3P_{\widehat{N}}(m) \\ \eta 3P_{\widehat{N}}(m) & : \quad \text{otherwise} \end{cases}$$

where $P_{\widehat{S}}(m)$, $P_Y(m)$, $P_{\widehat{N}}(m)$, are the power spectrum of the estimated clean speech, noisy speech, and estimated noise respectively with the percentile coefficient $\eta = 0.008$.

The square root of the resultant power spectrum $P_{\widehat{S}}(m)$ was taken and then recombined with the phase information, to give the enhanced speech.

4.3.3 Pollák, Sovka, Uhliř

Pollák *et al.* presented a SS approach for the reduction of noise in a car in [105]. This work was interesting in that a directional reference microphone was used to sample the environmental noise. This allowed the estimate to be closer to the noise, which would corrupt the speech, and this was similar to the reasoning for the microphone placement shown in figure 4.1

The noise estimate in [105] was based on energy tracking, to determine speech/pause sections:

$$E_y = \sum_{n=0}^{N-1} |Y(m)|^2 \tag{4.8}$$

$$E_N = \sum_{n=0}^{N-1} |N(m)|^2 \tag{4.9}$$

where E_y, E_N, n, N, are the energy of the corrupted speech, and noise frames, the sample number and number of samples in the frame respectively.

From equation (4.9), a threshold was calculated, which determines the energy level above which the frame being examined was said to contain speech, and under which the frame was said to contain noise. In [105] this was set at being:

$$E_{thresh} = 1.5 E_N \tag{4.10}$$

An update to the noise estimate only happens when E_y breaks this threshold value, and the update is calculated using equation (4.3). The SS was magnitude based, using full wave rectification as shown in equation (3.1) Pollák *et al* do note though that this system worked well only for positive SNRs, and this was the reason for the introduction of the directional reference microphone.

4.3.4 Kang and Fransen

The final third party approach looked at was Kang and Fransen [106], which was used to improve the quality of linear predictive coded (LPC) processed noisy speech. The SS was the same as equation (4.7) with $\alpha = 2, \beta = 1$. In [106] it was suggested through examination of the history of speech energy below 1kHz, that only activity below this point need be examined, to determine speech/pause boundaries. They postulated that the activity of voiced speech in this range was significant, as voiced speech has a large first formant value, whereas the resonant frequencies of noise tended not to fall in that vicinity. This discrepancy it was claimed, allowed the activity in such a band to determine speech activity.

An example of this can be shown in figure 4.7 which shows how the energy below 1kHz changes, with transitions between speech and pauses. At 1.6 seconds a section of speech can be clearly seen, as well as a pause section at 8 seconds onwards.

This calculation was made for each frame of the corrupted speech, and the maximum and minimum values in each frame were noted. This information was combined with the results for previous frames, to determine the overall maximum and minimum energy values for the

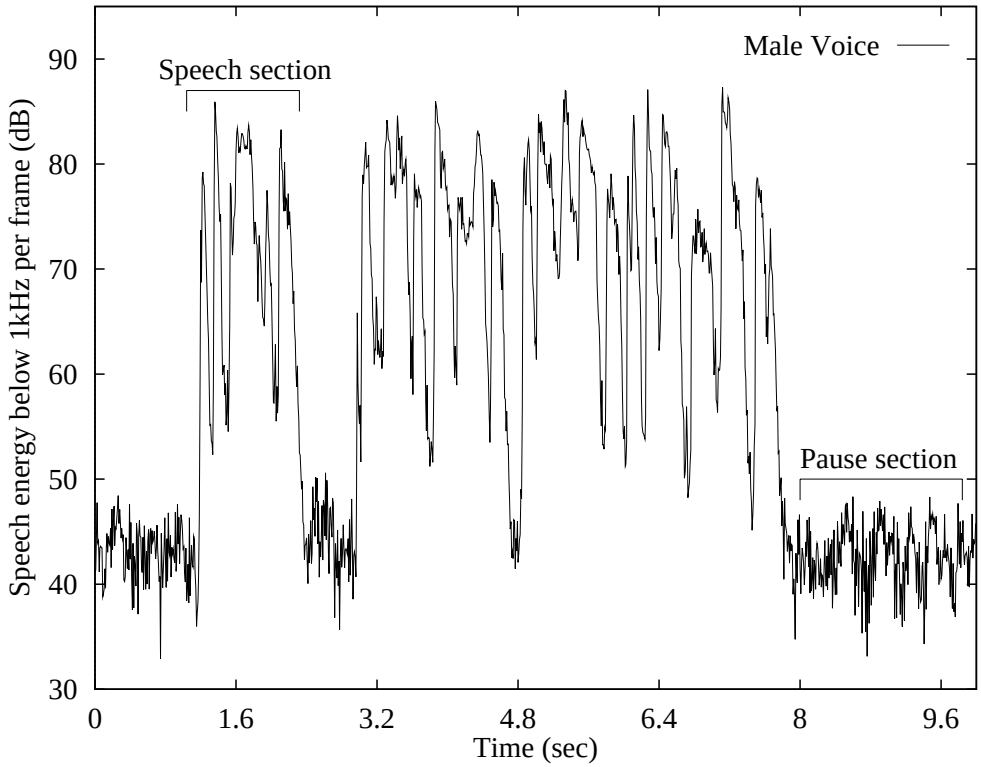


Figure 4.7: *Speech energy below 1kHz*

signal, and a threshold was set. When the low band energy of the current frame P in [106] was below this threshold level, then the signal was noise only and an update could be performed. The threshold was calculated by equation (4.11), and the update performed by equation (4.12)

$$Thresh = MIN + \frac{MAX - MIN}{8} \tag{4.11}$$

$$P_{\widehat{N}}^{new}(m) = p(P_{\widehat{N}}^{old}(m)) + (1-p)(P_{Y_m}) \tag{4.12}$$

where p is the feedback factor.

At first glance, it seems that this technique may be biased against unvoiced speech, which with its noise-like nature and potentially lower energy value (compared to voiced speech) may be unfairly categorised as noise causing updates to be performed then as well. Kang and Fransen claimed this effect was not severe in their algorithm, as they believed that unvoiced speech portions were brief in comparison to the length of speech pauses. The feedback factor p in [106], was set at $\frac{15}{16}$. In [106] this technique was tested with a variety of noise conditions from a quiet

room, to tank noise, and the inside of a naval destroyer information room.

The final method used in the PANE technique was the NEAH technique as described in section 4.2.

4.3.5 Voting Decision

The motivation behind the PANE technique was to apply a number of noise estimation techniques at once, to obtain better noise estimation performance. Using this, it was hoped that the potential for false detection of speech as noise could be reduced. To enable this, a voting decision process was implemented, which examines whether the individual branches detected a frame of noise. This is shown in figure 4.8.

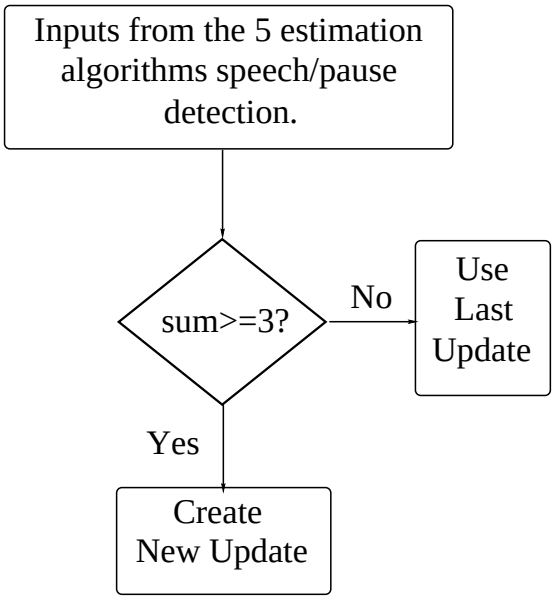


Figure 4.8: *The voting decision process*

Each estimation technique output a 1, if a frame of noise had been found, or a 0 if a frame of speech had been found. These were then added together to give a sum value, which could range between $0 \rightarrow 5$. As there were 5 techniques, it was decided that a frame of noise was detected if three of the five detection techniques agreed i.e. a majority voting decision. This allowed some leeway for techniques which may erroneously detect speech as noise, and vice versa.

4.3.6 Noise update method

In figure 4.8, it could be seen that once the PANE technique detected a frame of noise, an update occurred. The final process in the PANE technique was the decision on how the noise update should be performed. It was decided, that the information available for the update, should only come from those branches of the parallel realisation which had detected a frame of noise. An averaging process of the sets of values was rejected, due to the inaccuracy which it would introduce. The other options were to take the minimum or maximum of the values presented.

$$\widehat{N}_n(m) = \min\left(\widehat{N}_n^1(m), \widehat{N}_n^2(m), \widehat{N}_n^3(m)\right) \quad for\ n = 1, 2..128 \tag{4.13}$$

$$\widehat{N}_n(m) = \max\left(\widehat{N}_n^1(m), \widehat{N}_n^2(m), \widehat{N}_n^3(m)\right) \tag{4.14}$$

where $\widehat{N}_n^1(m)$, $\widehat{N}_n^2(m)$, $\widehat{N}_n^3(m)$ represent an example, when three of the five noise estimation techniques have detected a frame of noise.

Tests were performed with a variety of different noise sources and SNRs, to determine which of these two would provide better performance over the range of the signals of interest. A sample of two of these is shown in figure 4.9, and the rest of the experimentation is shown in Appendix C.

Figure 4.9(a)(c) shows the effects of taking the MIN and 4.9 (b)(d) MAX values. Looking at these, and the spectrograms shown in figures C.1 to C.4 it can be seen that over the range of SNRs and corruption signals used, the amount of noise which was left in the MIN cases was still significant. Visually this is shown best in C.1(a) and C.2(a). Comparing each case to its corresponding MAX result it can be seen that this was a result which held true across all the conditions examined.

The reason for this is that the MIN approach tends to favour the estimation technique by Boll, which takes the average value across the frequency bins, rather than drawing estimates from each technique when required. The average value as previously stated will not be as perceptually accurate, and can cause not only poor noise reduction, but possibly have a detrimental effect on the speech being enhanced. The spectrogram results tend to agree with this. On the other hand, the results of figures C.1 to C.4 clearly showed that when the PANE technique chose the MAX estimation sample from the techniques, it provided a more acceptable output in terms

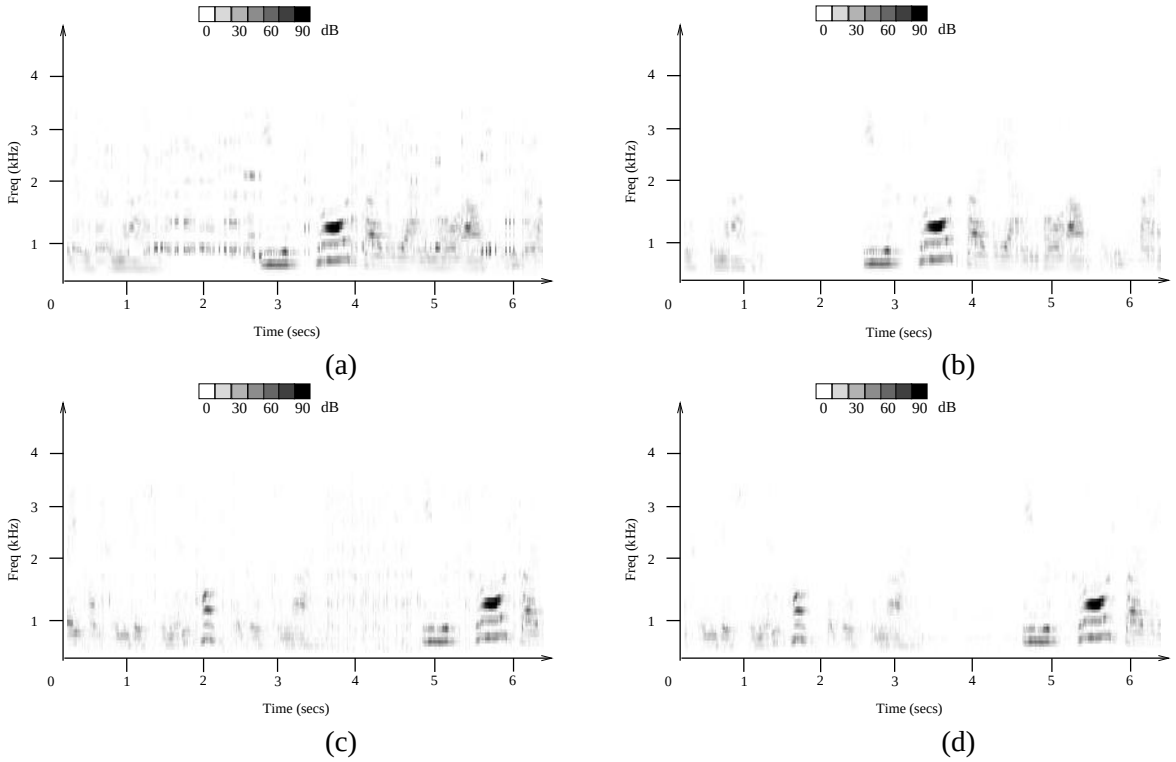


Figure 4.9: *Spectrograms sample of PANE technique with (a) Female/Music* 0*dB with MIN decision (b) MAX decision (c) Male/*1*kHz tone* 0*dB with MIN decision (d) MAX decision.*

of noise estimation. This was clear from the lack of corrupting noise present in the residual speech, compared to the MIN case, suggesting that this was a method which was robust over a range of SNRs and noise conditions.

4.4 Performance of the NEAH and PANE techniques

Although the preliminary tests of the NEAH and PANE techniques gave encouraging results, it was important to establish the validity of the techniques, with respect to other noise estimation and reduction work. In order to do this a series of statistical and perceptual tests were performed on NEAH and PANE, along with the other techniques which make up the PANE approach. The first test was to try and establish the amount of noise reduction which each technique could achieve. To simulate the conditions that a mobile phone user could experience two types of corrupting noise were used: pink noise, as it is considered to be a good estimate of the conditions present in general environmental noise; or a competing male speaker, which could simulate crowd noise or tannoy announcements.

4.4.1 SNR improvement results

The primary speech used was from a female speaking the phrase *"A man attacked by a leopard is in a stable condition in hospital tonight. He was clawed by the animal after he climbed over a safety fence at Marwell Zoo near Winchester to..'.* The corrupting male speech was *"A system error has occurred. Please hang up and try again later, or dial one nine two for the full directory enquiries service."* The female speech was mixed with one of these noise signals at four different SNRs ($-10,-5,0,5$ dB), and the noise estimation techniques run to determine the improvement in SNR. The results of this experimentation are shown in figures 4.10 and 4.11.

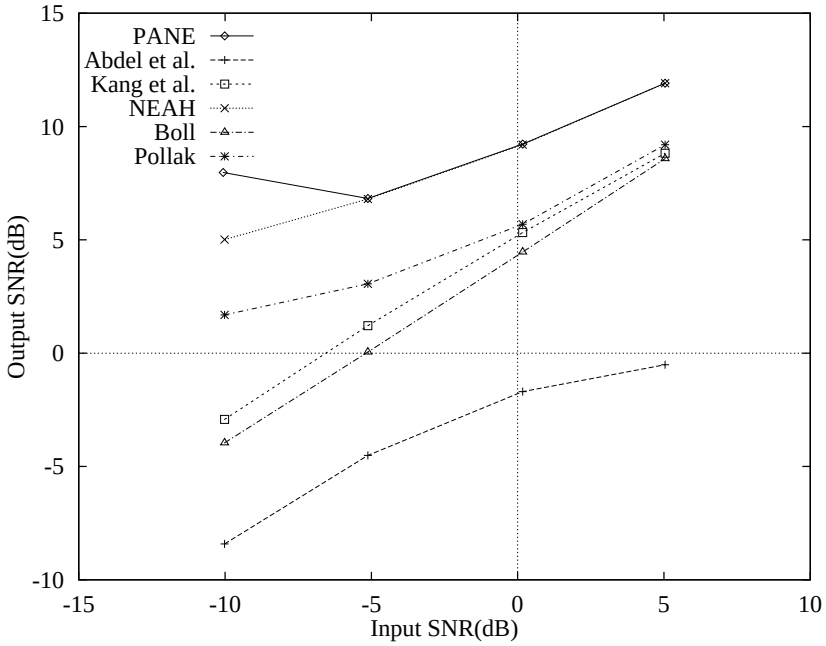


Figure 4.10: *Pink Noise SNR experiment results.*

It should be noted that the SS case used in these experiments was magnitude SS, as developed in Chapter 3. This was not the optimal SS for the estimation techniques of Kang *et al.* and Abdel *et al.*, which were originally utilised in an algorithm using power SS. To ensure a fair comparison between all the techniques was made, an extra experiment using the pink noise case and power SS for these algorithms was performed. This is shown in figure 4.12.

In figure 4.10, it can be seen that as expected the output SNR improves for each individual noise estimation technique as the input SNR increases for the pink noise corruption. The similar performances of the Kang *et al.* and Pollak *et al.* approaches was not unexpected, as they approach the estimation in a similar way. They only differ in the very low SNR regions, leading to the belief that Kang's approach was not as useful in such conditions. This made sense as

at low SNR the low band energy used for detection of speech/pause sections would be heavily corrupted by noise, making it more difficult to determine which sections were noise only. Boll's algorithm was not far behind these two in terms of SNR improvement, which was a surprise given that it was by far the most simplistic approach, consisting of only frequency averaging. The NEAH and PANE techniques though, outperform any of the other algorithms, with at least a 3 to 4 dB improvement of the PANE over the NEAH technique at very low SNRs, and as much as 5dB more than the next nearest system. Noticeably though there was a drop in the results for the parallel case between -10dB and -5dB input SNR and this was caused by the use of the multiple techniques in PANE. As each routine in PANE constructs its own noise estimate in a different way, the errors presented by each will differ. This is shown by the spectrograms in Appendix C and 4.9. The NEAH technique performs almost identically to the PANE in terms of SNR, apart from -10dB input SNR. It seems that the parallel architecture of the PANE contributes to greater noise suppression at -10dB, but the technique from then tracks the NEAH results. This was almost akin to a diversity gain in the way which the technique averages out any errors that may occur. The worst performing algorithm was that proposed by Abdel *et al.*, but this was to be expected as this algorithm relies on the use of power SS instead of magnitude SS, meaning that the testing was not reflecting this technique in its original form.

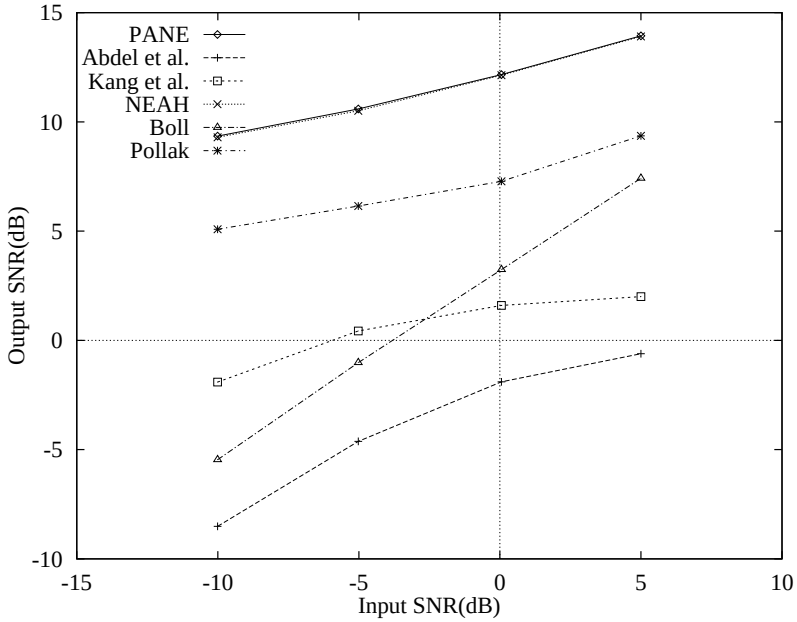


Figure 4.11: *Competing Male Speaker SNR experiment results.*

Figure 4.11 shows the results of the same experimentation using the Male voice as the corruption, instead of the pink noise. Again, it can be seen that the NEAH and PANE technique

perform better than any of the other algorithms, in terms of SNR improvement. Interestingly up to 0dB, the PANE technique was marginally the best performer. The NEAH technique was the next closest performer, but it can be seen that the PANE draws on its parallel estimation to obtain marginally better SNR improvement. Again the Abdel *et al.* and Kang *et al.* have some of the lowest improvements due to their reliance on power SS. There was no dip in performance of the PANE technique, which was shown in the pink noise case. As the interference was a competing speaker, and the operation was on small time frames to make the primary speech stationary, by default the interfering speech also was stationary, within the boundaries of the frame.

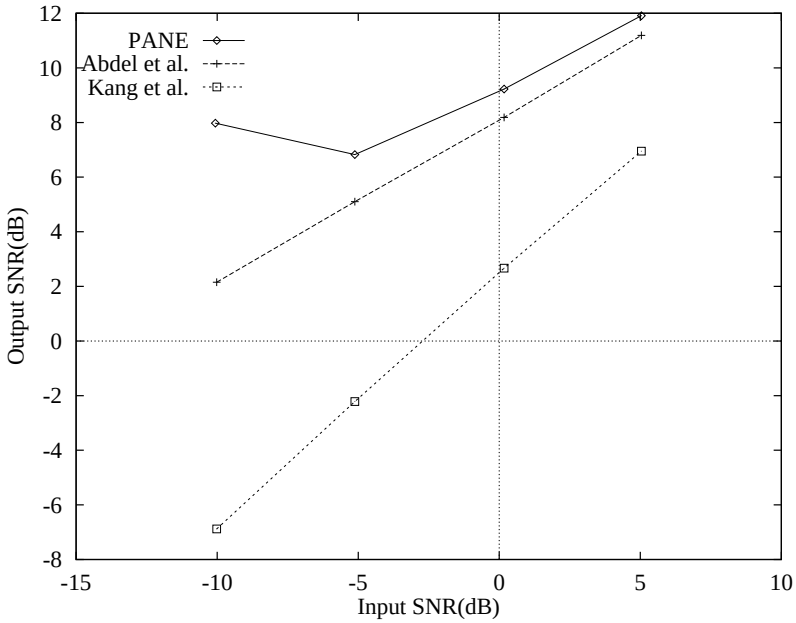


Figure 4.12: *Pink Noise power SS SNR experiment results.*

Figure 4.12 shows the algorithms which use power SS in their implementation rather than the magnitude SS, which the PANE and NEAH techniques used. It should be noted though that in these results the PANE technique was still using magnitude SS, and is placed here for a comparison. The Abdel *et al.* algorithm improves significantly by moving to power SS, improving by up to 11dB. The Kang *et al.* algorithm still struggles with the low SNR cases for the reasons earlier stated. The PANE technique though still outperformed these, even when they were running as originally intended.

4.4.2 Listening test results

While the SNR results gave a good indication as to the amount of noise reduction which each algorithm could provide, it gave little idea as to how the resultant speech actually sounds. The only way to to determine the quality of the speech each noise estimation technique produced in terms of both noise reduction and speech corruption, were subjective listening tests. Using a cross section of listeners, a measure of the respective accuracy of each algorithm could be obtained.

A group of 10 subjects were exposed to eight test sequences each. The test signals are:

1. Male Speech/Pink Noise -10dB SNR
2. Female Speech/Male Speech -10dB SNR
3. Male Speech/Female Speech -5dB SNR
4. Female Speech/Pink Noise -5dB SNR
5. Male Speech/1kHz Tone 0dB SNR
6. Female Speech/Music 0dB SNR
7. Male Speech/Music 5dB SNR
8. Female Speech/1kHz Tone 5dB SNR

Each test signal was processed by the NEAH, PANE, Boll, Abdel *et al.*, Kang *et al.* and Pollák techniques, and the resultant speech of each recorded onto DAT. Using a DAT machine and a pair of Sennheiser HD25SP headphones, the subjects listened to the 8 sets of speech outputs, and marked them according to the Mean Opinion Score (MOS).

The MOS is one of the most widely used subjective quality measures, and using it, listeners rate the speech on a five point scale, where the subjective impression of the listener is represented by a numerical value. It is common for the listeners to be presented with some form of training information, to which they relate the test information. In this case before each set of test results, examples of the clean and corrupted speech were presented, which represented the best possible (5 on the MOS scale) and worst possible (1 on the MOS scale) speech which could occur. After this, the listeners were presented with the output of the techniques. In order to prevent any subjective biasing of the results, the order of the techniques were randomised from test to test.

Each listener was presented with the following guidelines:

> The purpose of the test is to determine how well the algorithms remove the interference from the speech. On the tape, you are hear 8 sections of speech. The first two give an indication of good clean speech, and the worst possible corrupted speech. These are to allow you to judge the performance of each algorithm.
>
> After the first two sections of speech, you are presented with 6 segments from different enhancement algorithms. In the table below please enter a score for each segment as indicated on the following scale;

Rating	**Speech Quality**	**Level of Distortion**
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

> The qualities you should keep in mind as you judge each segment is;
> 1. How good the Speech quality is.
> 2. How much does the distortion affect speech quality.

The advantage of the use of MOS testing was that each listener was able to determine what they felt were "good" results, and allowed a great deal of flexibility for the listener to judge as they perceive. The disadvantage of such testing, was that some listeners may have a more discriminatory hearing, and may be more critical than listeners who do not. To try and eliminate this, one pair of so called "golden ears" were used, to help prepare the tests and then a random selection of 10 listeners were chosen to perform the tests. In order to keep the test conditions consistent, a quiet office was chosen to locate the test devices. The results of the MOS testing are found in figures 4.13 and 4.14, having been split between two graphs, for easier examination.

Each plot on figures 4.13 and 4.14 show how the average MOS score varied between algorithms for each test signal. It was clear that in general the NEAH and PANE techniques provided higher MOS scores (and subjectively better quality output speech) than any of the other techniques. It was interesting to note though, the perceived change in performance of each technique as the test signals and SNRs changed. For example, the Abdel *et al.* algorithm was perceived to work well for the test condition male voice/1kHz tone at 0dB, but poorly for male voice/pink noise at -10dB SNR.

Of the other approaches presented, that by Abdel *et al.* gave the closest performance to the developed NEAH and PANE techniques, with the energy tracking method of Pollák *et al.* provid-

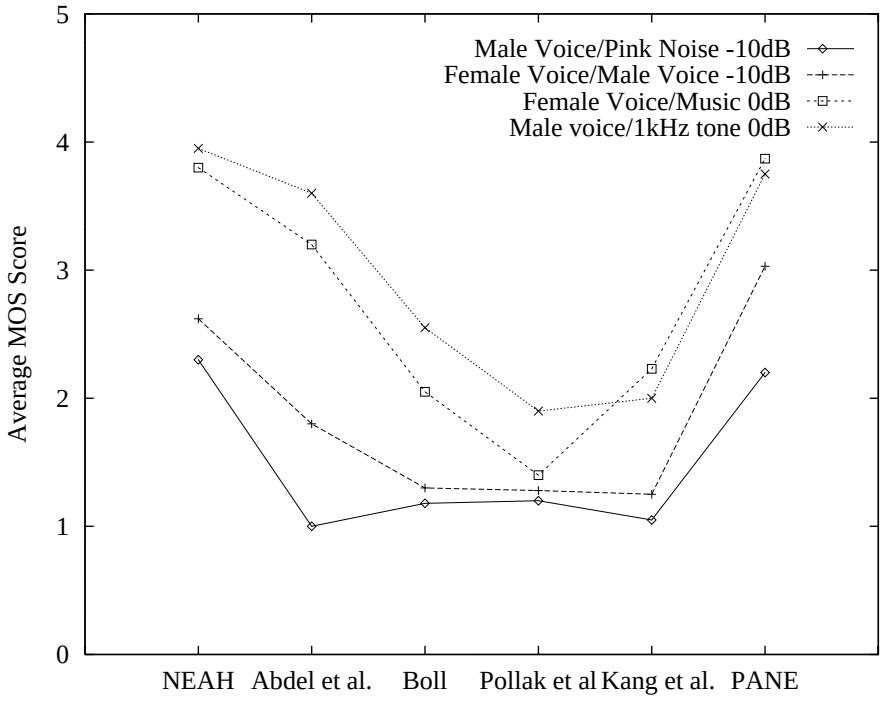


Figure 4.13: *First set of MOS listening test results.*

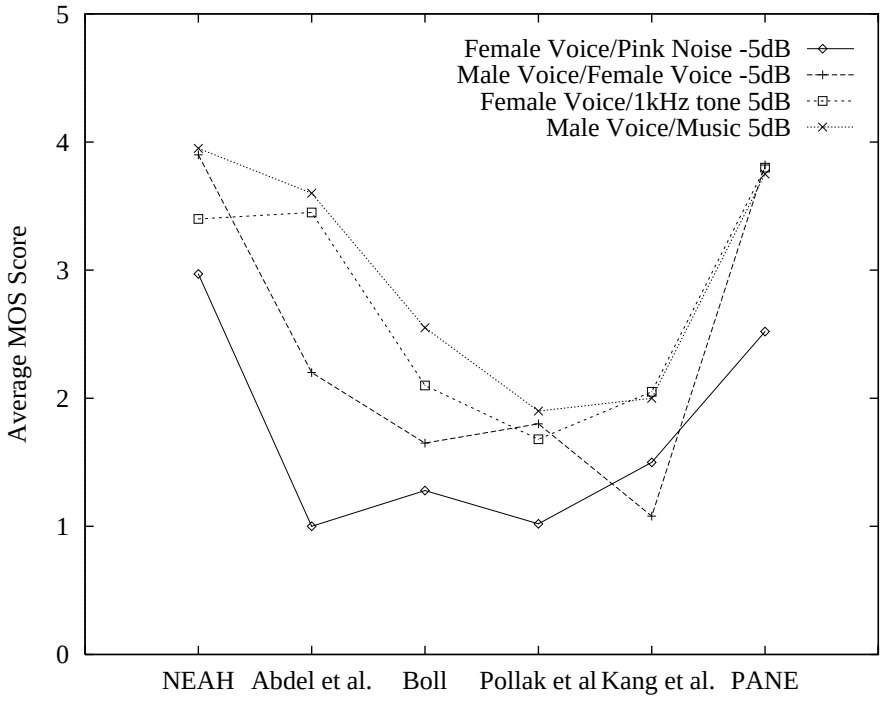


Figure 4.14: *Second set of MOS listening test results.*

ing the worst perceived output. Comparing the MOS scores for the NEAH and PANE techniques, it was interesting to note that in figure 4.13, apart from the test case male voice/pink noise at 0dB, the PANE technique is considered to provide better quality speech. This trend is reflected in figure 4.14 where again apart from one case (Female voice/Pink Noise at 5dB) the PANE technique was thought to provide superior speech quality. Looking closely though, it can be seen that for each algorithm in general the MOS scores increased as SNR increased. It can still be seen though that even for the NEAH and PANE techniques, the MOS results in some cases varied between poor (2) and fair (3), for cases with very low SNR. At a first glance, this may be seen to be a drawback of the algorithm, although looking at the situations where these scores occurred, it can be seen that the SNRs are negative (either -5 or -10dB), which were conditions unlikely to be experienced by a mobile phone user. These results have been included though to show how the algorithm would perform in such cases. Those results relating to 0 dB and above, gave a closer indication of conditions the mobile user may experience, and in these the NEAH and PANE techniques produced speech of a good quality (as denoted by the MOS score).

Looking back at the MOS test definitions, it could be seen that not only was the quality of the speech an issue, but also the level of distortion of the speech. The MOS test definitions rate the intrusion of this from, *Very annoying and objectionable* to *Imperceptible*. Even with the NEAH and PANE techniques, the best level of distortion which was obtained was *Just perceptible but not annoying*, meaning that the speech enhancement algorithm as it stood introduced a small amount of distortion into the speech. This was one of the tradeoffs of standard spectral subtraction approaches. When noise reduction increased there was a corresponding increase in the distortion of the speech signal. In Chapter 5 an approach to try and combat this distortion is examined.

The results of the MOS testing were important, as while the SNR graphs of figures 4.10 and 4.11 showed almost identical results for the NEAH and PANE techniques, the subjective testing showed that each was perceived, to give a different quality output. This enforced that, in speech processing results that may look similar statistically, may differ in how they are subjectively perceived.

4.5 Summary

In this Chapter, the criterion for an accurate noise estimation technique was discussed, along with the assumptions which were made for estimation. Single and two sensor approaches were examined, and a two sensor approach chosen to allow tracking of the background noise, even during speech utterances. The NEAH technique was proposed, which looked to reduce inaccuracies in the estimation process, by verifying whether a section of signal was speech or noise, before conducting an estimate update. The idea of improving the estimation process by applying multiple noise estimations in parallel was discussed, and the PANE technique suggested. Each of the techniques which were used in the PANE were examined, along with the rest of the enhancement algorithm they operated in. It was shown through the use of spectrograms, that the PANE technique operated better in a non-stationary environment by use of a MAX decision process after the voting decision. This allowed the PANE technique to draw on the best parts of each of the constituent algorithms.

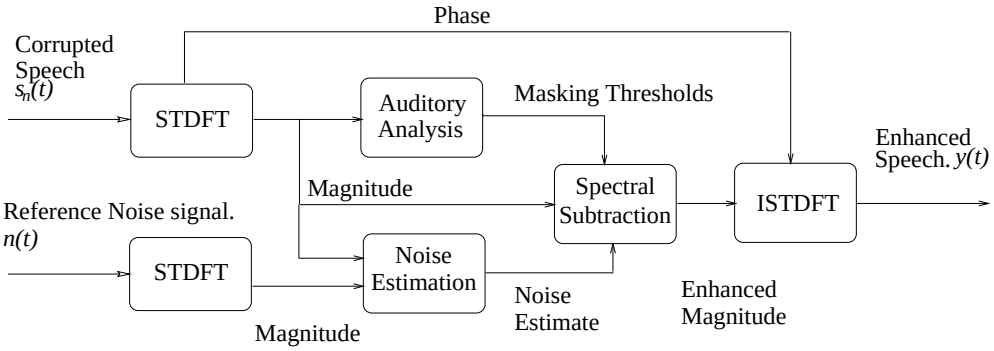


Figure 4.15: *Flow Diagram of the system with the noise estimation included.*

Through the use of statistical and subjective testing, the NEAH and PANE enhanced speech were compared to the other methods examined. The SNR results showed that the NEAH and PANE techniques substantially reduced the level of corrupting noise in the speech signals, with the improvements ranging from $5 \rightarrow 15$dB. While promising, it was shown that SNR results do not tell the whole story. and a set of subjective listening tests were performed using the MOS. From these it was shown that the NEAH and PANE techniques produced *"good"* quality speech (as defined by the MOS scale), although the PANE technique was perceived to provide slightly better quality speech. Due to this, the PANE method was taken as the estimation algorithm for the speech enhancement technique being developed. Following the integration of the noise estimation into the speech enhancement algorithm, the process is now shown in figure 4.15

Chapter 5
The Removal of "*musical noise*" from Enhanced Speech

In this Chapter, the phenomenon of musical noise is examined. A new technique is proposed, to attempt to reduce the level of musical noise that corrupts speech, enhanced by spectral subtraction. By examination of the spectral properties of the enhanced speech, points are defined which could potentially result from musical noise, and a perceptually acceptable replacement is substituted.

5.1 Effect of musical noise

In section 2.7, the topic of musical noise and its causes were introduced. Depending on the accuracy of the noise estimate, and the random removal of speech components by spectral subtraction, this *warbling* effect can be audible in the enhanced speech. In section 3.2.3 it was also mentioned that in the spectral subtraction algorithm, a method which could reduce the introduction of these tones, was the application of a spectral noise floor. The disadvantage of such a method was the degradation of the overall noise reduction property of the system. An example of the effect of musical noise on enhanced speech can be seen in figure 5.1(b), where spectral points have been introduced which are not part of the original speech shown in 5.1(a). There are also a number of speech points which have also been removed. Both of these conditions lead to the musical noise phenomenon.

Along with the application of a noise floor, another technique which can be implemented as part of SS in an attempt to reduce the appearance of musical noise, is oversubtraction of the estimated background noise. This approach was examined in section 3.2.2, as part of the SS algorithm, but it was felt that using this led to an increase in the removal of the desired speech signal, which was more noticeable than the musical noise, (as shown in the results of section 3.2.2). These techniques were designed to reduce the amount of musical noise which could

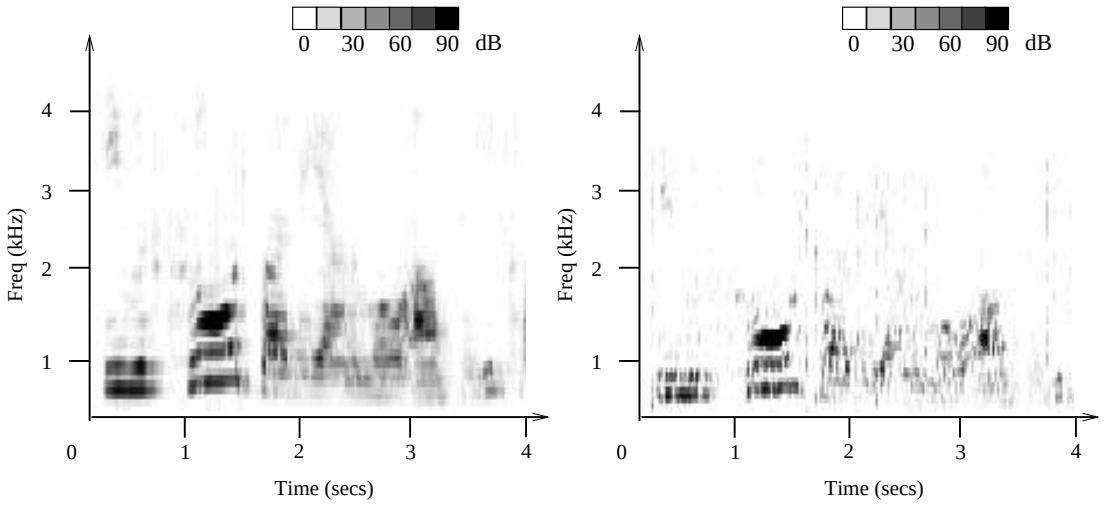


Figure 5.1: *Spectrogram of (a) Clean Female Voice (b) Musical Noise corrupted Female Voice*

occur in a signal, where as the problem being examined, was how to deal with any musical noise which did occur.

It was therefore reasonable to attempt to operate only on the portions of the signal where the effect was most perceptually disturbing, rather than dealing with the signal as a whole, which the oversubtraction did. At this point, it should be noted that the small portion of noise floor shown in section 3.2.3 was still used, as this helped give a output which was more pleasing to the ear, without affecting the SNR of the output signal adversely. As the tradeoff in SS places noise reduction against speech distortion, it was felt that a separate technique which could address the problem of musical noise in the enhanced speech would be beneficial. This would have the advantage of being able to operate on a variety of conditions, and be separate from the SS process.

The difficulty with musical noise is predicting where it affects the enhanced speech and audio. This relies not only on the accuracy of the noise estimate, but the SS and the masking of the speech by noise and vice versa. The effect is random, with no two test conditions exhibiting the same temporal/spectral positions, as well as the same degree. In this respect, it was felt that it was more beneficial to approach the problem from an auditory viewpoint, rather than a strictly statistical basis. By examination of the spectral content of the signal, it may be possible to adapt to the condition the technique was applied to, as well as allow the introduction of the masking thresholds, to determine whether the replacement values were perceptually valid.

5.2 How the problem of musical noise has been approached

As stated in the previous section, the presence of musical noise was audibly apparent, due to the warbling sound which it overlays on the enhanced speech. Identifying the points which contribute to this audible affect, required the ability to discriminate between speech sections and musical artefacts. In [97] musical noise was described as having a lower mean power but a larger variance than the original noise.

This meant, that any point which could be categorised as musical noise would exhibit a high variance with regard to the surrounding spectrum. This could be used as a classifier, by examining each spectral point and the variance of the area surrounding it. To do this would be time intensive. If the technique could pick out potential points, then use this classification to determine musical noise points from sections which were isolated speech segments, then the computational time could be reduced. A second classification method was described in [107], which stated that after SS has been applied, the short time magnitude spectrum can contain a succession of randomly spaced spectral peaks, in between which the spectral points are strongly attenuated.

This suggested that the presence of tonal artefacts could be detected, by examining the size of the spectral point, compared to those surrounding it in time and frequency. This classification process was also suggested in [108]:

> "After the noise PSD estimate is subtracted from the input signal PSD, zeros are placed in the resulting PSD at discrete frequency locations that have abnormally large noise power estimates..."

In this case, [108] defined *abnormally* as a power level of four times the average over all frequencies. This led to the supposition that musical noise points could be detected by:

1. A larger magnitude value than the surrounding spectrum points.
2. A larger variance than the surrounding spectrum.

Both of the conditions rely on examination of the spectrogram, as described by [109] for variance classification.

5.2.1 The work of Whipple

In [109], Whipple proposed a classification method, which operated in both time and frequency across the spectrogram, to identify possible musical noise points. The process examined several frames proceeding and following the frame containing the point of interest (POI). It was reasoned, that spectral peaks which belonged to musical noise would tend to be isolated in both time and frequency, causing them to stand out in the spectrogram, compared to the surrounding spectrum. On the other hand, spectral peaks which occupy several consecutive frames of the signal were more likely to be due to speech.

The analysis was performed by defining two regions, as shown in figure 5.2, where the POI(s, t) lies in region B. The POI(s, t) lies at sample number s in window number t.

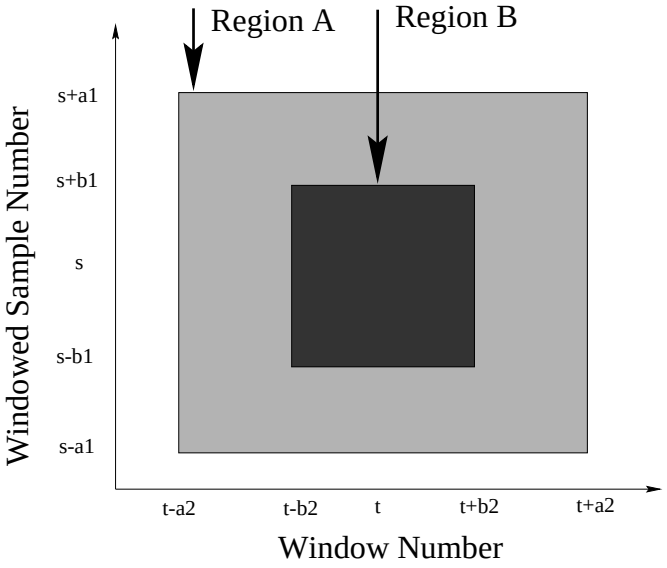


Figure 5.2: *The analysis regions as defined by G Whipple*

The height and width of the regions A and B around POI(s, t) were defined by the parameters $a1,\ a2,\ b1,\ b2$ which in [109] were set to $a1 = 3,\ a2 = 3,\ b1 = 2,\ b2 = 2$. This defined a 5×5 square around POI(s, t) for region B and a 7×7 square for region A. The decision as to whether the POI(s, t) was a speech or noise point, was made by examining the ratio of the energies contained in regions A and B, using the following:

$$E_B(r,t) = \sum_{i=t-b2}^{t+b2} \sum_{m=r-b1}^{r+b1} X_i(m) \tag{5.1}$$

$$E_A(r,t) = \sum_{i=t-a2}^{t+a2} \sum_{m=r-a1}^{r+a1} X_i(m) - E_B(r,t) \tag{5.2}$$

where $E_A(r,t)$, $E_B(r,t)$, $X_i(m)$, r, t, i are the total energy of region A, region B, the magnitude spectrum, the points defining the centers of regions A and B, and the frame number respectively.

if $E_B(r,t) \geq \gamma E_A(r,t)$
Isolated peak probably due to musical noise.

if $E_B(r,t) < \gamma E_A(r,t)$
Does not contain an isolated peak and is probably a speech point.

In [109], γ was defined as having a value of ten. If the region B was found to have an isolated peak, the magnitude spectrum which were contained in that region were set to zero to remove it.

$$X_i(m) = 0 \tag{5.3}$$

If region B was found not to contain an isolated spectral peak, then the magnitude spectrum of that region were left untouched. An example of the performance of this technique can be found in figure 5.3.

It can be seen in figure 5.3(d) that the technique as described in [109] was able to remove some of the tonal artefacts which have been left in by spectral subtraction alone. Unfortunately, when compared to figure 5.3(a), it can also be seen that some of the required speech portions had also been lost. This was due to the methodology of replacing all the spectral points in region B with zeros, which had the possibility of removing speech sections close to musical noise artefacts in the signal.

Of more importance, figure 5.3 showed, by examining the peaks of the magnitude spectrum, it was possible to determine points which could be musical noise, as opposed to speech segments. The disadvantage of erroneously removing some of the speech spectrum needed to be addressed, for such a technique to be useful. The use of a second classification technique may allow the detection of portions of speech which had passed through the energy criterion, but

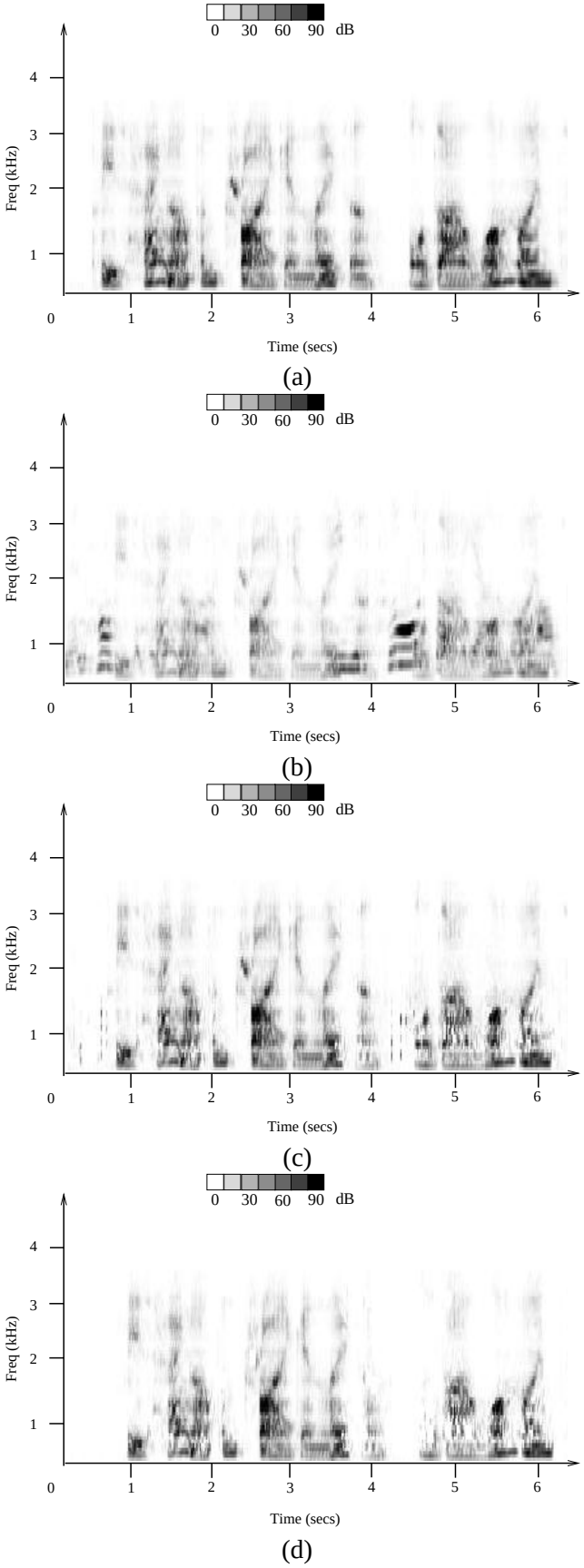


Figure 5.3: *(a)Clean male speech (b) With corrupting female speech (c) After SS alone (d) After SS and Whipple's technique*

should not be removed, while ensuring that musical noise artefacts were detected. The second element which could be exploited for this is the variance of the spectrum round about the POI.

5.2.2 The Work of Z Goh, KC Tan and BTG Tan

In [110], a musical noise suppression technique was proposed, which attempted to detect musical noise points, by examining the variance of the region around the POI. This was motivated by the belief that the work of [6] and [109] cannot remove musical noise which appeared not only as isolated peaks, but *short ridges* in the spectrogram. As stated in the previous section, a more rigorous detection method was required.

Goh *et al.* theorised that musical noise generally affected only those portions of speech which had low energy (weak unvoiced speech). As unvoiced speech is accepted to be noise-like, it was these portions of speech which were most likely to be corrupted, and removed during enhancement. Strong voiced speech segments contain enough energy to pass thorough energy classification, such as in [109], or by the common noise oversubtraction method. In order to prevent the loss of these points, while still ensuring that musical noise was suppressed the following technique was proposed.

The classification of speech points, as opposed to musical noise or speech with low energy, was obtained through the oversubtraction method. Having obtained the spectrogram of the desired signal from SS with a large β value (16 in [110]), the points were classified by:

$$\begin{aligned} &\text{if} \quad \widehat{S}_i(m) > 0 \quad \text{Speech point. (Region 1).} \\ &\text{if} \quad \widehat{S}_i(m) < 0 \quad \text{Anomolous point. (Region 2).} \end{aligned}$$

where $\widehat{S}_i(m)$ is the magnitude spectrum of the enhanced speech for frame number i.

There was a possibility for the magnitude spectrum to contain negative values, due to the oversubtraction factor used in the SS. If a spectral point was classified into region 1 then it was a high energy speech point, and was strong enough to be retained in the spectrogram. Any points which were classified into region 2, could potentially be classified as musical noise, and more classification must be performed to ensure that the voiced speech points were not altered, leading to a drop in the intelligibility of the speech. As energy classification cannot be used in this case, the region surrounding the POI was covered by six blades, which intersected the POI, as shown in figure 5.4.

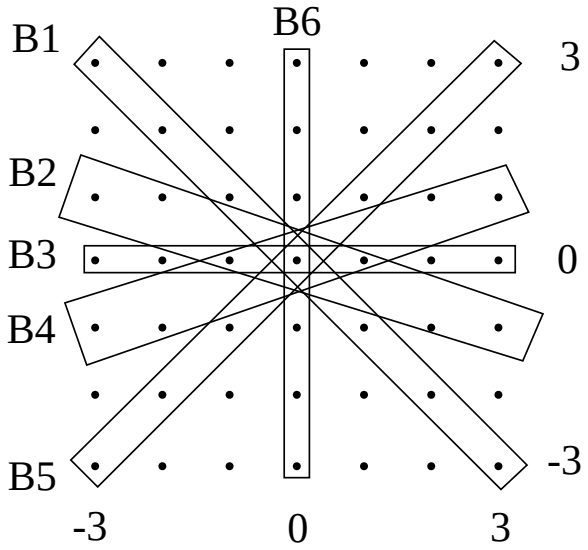


Figure 5.4: *Six blades over a section of* 7×7 *spectrogram points round about the example point (*0,0*)*

The width of the blades was chosen, to ensure that the grid points are intersected in a straight line, and that an equal number of points were contained in each blade. The length was chosen, so that they were longer than most of the short stripes that may have occurred due to musical noise, but shorter than the lengths of strips which occurred due to speech signals. In order to determine whether one of the anomalous points was a weak speech point or noise, the variance of the spectrum values the blades intersect was calculated.

$$\mathrm{var}(B_c) = \sum_{(i,m)\in B_c} \frac{20\log_{10}(|\widehat{S}_i(m)|+1)^2}{|B_c|} - \left\{ \sum_{(i,m)\in B_c} \frac{20\log_{10}(|\widehat{S}_i(m)|+1)}{|B_c|} \right\}^2 \quad (5.4)$$

where $|\widehat{S}_i(m)|$, B_c, (i,m), c are defined in [110] as the enhanced speech magnitude, the blade being measured, the spectral co-ordinated of the POI as defined in time and frequency, and the blade number respectively.

Having calculated the variance of all the blades which intersected the POI, then the blade with the minimum variance B_{min} was identified. The variance of this blade $\mathrm{var}(B_{min})$ was used to indicate whether the POI was in fact a section of weak energy speech, or a musical noise point. For points which correspond to low energy speech then due to the collection of other speech spectral values around the POI, the variance would tend to be low. Points which correspond to

musical noise though would exhibit a great variance, as the blade length was longer than the musical noise point, and the other points in the blade would be filled with background spectral values. This would cause $var(B_{min})$ to be considerably higher, meaning that a threshold could be set, over which points could be classified, as definitely being musical noise.

$$(i,m) \in Region\ 1 \text{ if } var(B_{min}) < \tau \tag{5.5}$$

where τ is the threshold which separates musical noise points from weak energy speech points. This was set to $\tau = 200$ in [110].

These two layers of classification were designed to ensure that it would be unlikely that any speech points could be available for correction by the technique. The two stages work in tandem, as it was possible for there to be short burst of speech which have rapidly increasing magnitude values, giving rise to a high $var(B_{min})$. If only the variance calculation were applied as classification these points would be classified erroneously as musical noise. The two layer approach that was proposed in [110], ensured that sections of speech like this would be recognised under the magnitude classification.

Following the classification of the spectrogram into speech and probable musical noise sections, the processing of the points in region 2 was performed. In [110], it was assumed that even after this classification, it may be possible for low energy speech points to be erroneously classified as musical noise. In this case, the processing on the region 2 points must ensure that while genuine musical noise sections were severely attenuated, any speech which remained in this region must be altered as little as possible. In order to achieve this, [110] proposed the use of the median of the spectrogram points which the blade B_{min} intersected:

$$|\widehat{S}_i(m)| = median_{(i',m')\in B_{min}}(|\widehat{S}_i(m)|) \text{ if } median_{(i',m')\in B_{min}}(|\widehat{S}_i(m)|) < |\widehat{S}_r(k)| \tag{5.6}$$

Goh *et al* felt that using this form of replacement would result in any speech points being altered very slightly, due to the uniformity of the spectral points in speech strips. Conversely, the non uniformity of musical noise sections would lead to a replacement value which would be considerably attenuated, causing a reduction of the effect of musical noise.

Goh *et al* used this musical noise reduction technique on speech which had been enhanced with

SS $\alpha = 2$, $\beta = 1.8$. An example of the results of this technique are presented in figure 5.5, which shows the same conditions presented in figure 5.3. A comparison of figure 5.3(d) and figure 5.5(d) show that the computationally intensive method of Goh *et al* does cause less degradation of the speech, than that presented by Whipple. Comparison of these results does show though that while the method of [109] was computationally very simple, it was still able to remove a great deal of the musical noise from the enhanced speech signal. When it is taken into consideration that figures 5.3 and 5.5 represent a case where the SNR is -5dB, it shows that the energy based classification has a valid contribution to be made in a musical noise reduction technique.

While both of the techniques presented here do reduce the amount of music noise considerably, there was still a degree of speech removed from the signal, which can be seen in figure 5.3(d) and figure 5.5(d). The corruption was especially noticeable in the lower frequency ranges, where there are small section of speech removed. The goals behind the speech enhancement technique being developed were a reduction in the background noise and a maintenance of good speech quality. Clearly, while both the Goh *et al* and Whipple methods resulted in a drop in the musical noise, the quality of the speech can be seen to suffer. The difficulty was to determine how this could be achieved, as clearly both the methods for classification of the speech/musical noise still caused a corruption to the speech signal. This led to the idea of combining aspects of these two classifications, along with a method of replacing speech points which had been removed.

5.3 The developed musical noise reduction technique.

The examination of the methods presented in [109] and [110], showed that using the classification methods for musical noise, a great deal of the effects of this corruption could be reduced. The ability to counteract the removal of speech information was the prime area of concern, to extend the musical noise reduction into a non-stationary environment.

It was evident, that the variance classification of [109] was able to detect the musical noise points correctly. By more careful selection of the replacement value used for these points, the method could be improved. Goh *et al.* on the other hand developed an approach which was able to determine the difference between weak speech and musical noise points, although some corruption still occurred in the resultant speech. This method relied on the use of oversubtrac-

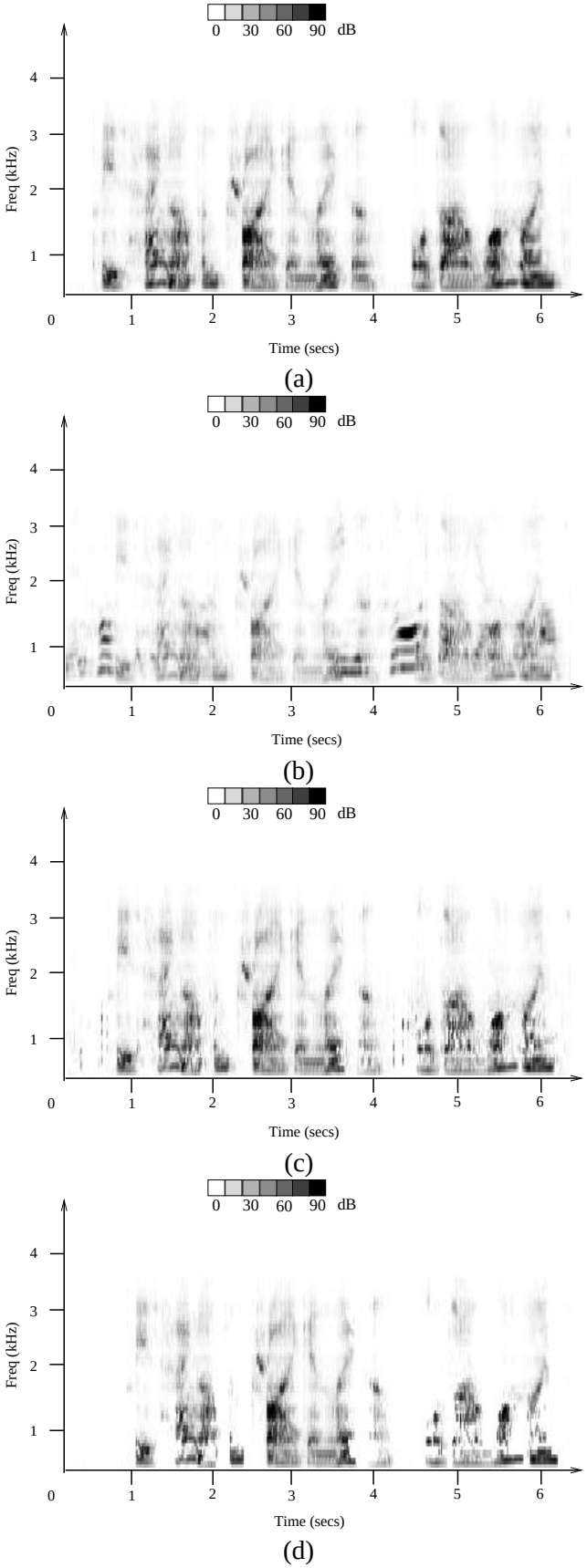


Figure 5.5: *(a)Clean male speech (b) With corrupting female speech (c) After SS alone (d) After SS and Goh et al’s technique*

tion, for the detection of potential musical noise points, which was an approach shown not to be successful for non-stationary noise in section 3.2.2.

If the successful classification method of [109], was attached to the front end of the variance calculation of [110], this could result in a drop in the computational time the technique will require. Not using the oversubtraction process exposes less weak speech to the classification process, as the subtracted noise signal was lower in value, resulting in less execution of the variance technique to separate the speech from noise. This also prevents some of the weak speech from being altered by the musical noise reduction technique, and should maintain the quality of the enhanced speech.

The full musical noise reduction technique is presented in figure 5.6. This shows the merging of the Whipple and Goh *et al.* work as well as the positions where the new sections have been implemented. Each part of the flow diagram shall be explained as the technique is described.

5.3.1 Power SS

In order to determine which points of the spectrogram were definitely speech, power SS was applied to the corrupted magnitude, and the estimated noise. This SS was not to obtain an estimate of the enhanced speech, but to determine which points could be subject to musical noise, and which were strong speech points that did not require to be operated on. Unlike the Goh *et al.* method though, there was no oversubtraction involved in the spectral subtraction. This resulted in the conditions:

$$\begin{aligned} &\text{if} \quad \widehat{S}_i(m) > 0 \quad \text{Speech point. (Region 1).} \\ &\text{if} \quad \widehat{S}_i(m) < 0 \quad \text{Anomolous point. (Possibly Region 2).} \end{aligned}$$

The threshold of zero was chosen, as the majority of the problems with musical noise rise from regions where negative values have arisen from the spectrum after SS. Using this operation, it was possible to classify those points which were definitely speech, from those that could potentially be musical noise. The potential musical noise points were marked in the spectrogram and only those were passed onto the energy classification, all the points in Region 1 were left untouched.

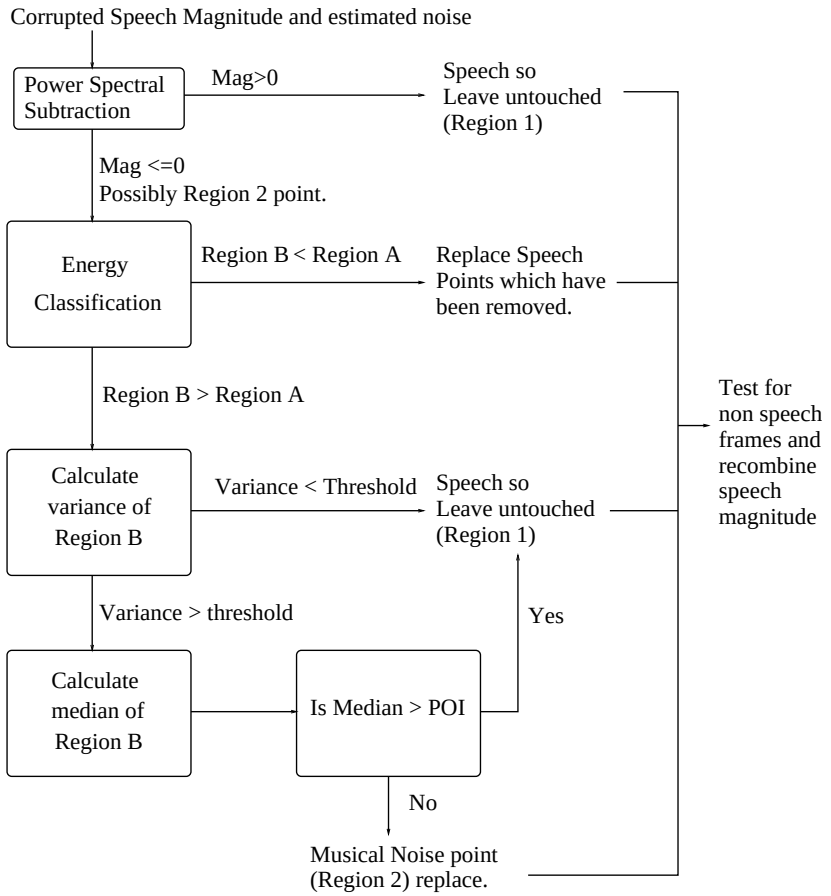


Figure 5.6: *Software flow diagram of the musical noise reduction technique*

5.3.2 Energy classification

Once the anomalous points were detected, they were passed to the first form of musical noise detection, the energy classification. From [109] and figure 5.2, it could be seen that about the POI a 5×5 square region for B and a 7×7 square for A were defined. It was felt this was excessive in terms of the computation required, and it was examined whether the use of a smaller 3×3 region B and a 5×5 would affect the operation of the classification. In order to do this, the Whipple technique was run, using these two sets of parameters, and it was assumed that all points outside the spectrogram were set to zero, to prevent wrapping around of the blade calculations.

Table 5.1 shows the average RMS percentage magnitude error results, with respect to the original speech from implementing the energy estimation as $7 \times 7/5 \times 5$ case as defined in [109], or as the proposed $5 \times 5/3 \times 3$ case.

	Percentage error (%)	
Speech operated on	$7 \times 7/5 \times 5$	$5 \times 5/3 \times 3$
male/pink noise -10dB	4.116969	3.648950
female/male voice -5dB	1.222826	0.867002
male/1khz tone 0dB	1.847490	1.434315
female/music 5dB	0.997472	0.647409

Table 5.1: *Average RMS percentage magnitude error results for changing region sizes.*

It can be seen that by using the proposed $5 \times 5/3 \times 3$ region sizes, that the Whipple technique by itself produced lower average RMS magnitude percentage errors. This was due to a smaller area of the spectrogram being analysed, and altered than that originally proposed by Whipple. This suggested that there was no penalty in detecting the musical noise points from analysing smaller areas in the energy classification. In this case the smaller region sizes were used which when looking at figure 5.2 gave parameters $a1 = 2,\ a2 = 2,\ b1 = 1,\ b2 = 1$, which defined the smaller regions.

The energies in regions A and B were summed as in equations (5.1) and (5.2), and from this classification, it was possible to separate points which were musical noise or speech, from sections of speech which were either retained or could have been removed. This could be seen in figure 5.2, where the sections of speech which were not detected as potential musical noise were passed onto a separate stage, replacing speech points which had been removed. This was possible, as any artificial nulls in the spectrum would result in the energy in region B becoming very small, compared to the surrounding speech points in region A. The energy comparison in this case used $\gamma = 1$, as it was required to more accurately separate out any potential musical noise points from speech which had made it through the SS classification.

if $E_B(i,m) \geq \gamma E_A(i,m)$
Isolated peak probably due to musical noise.

if $E_B(i,m) < \gamma E_A(i,m)$
Does not contain an isolated peak and, was probably speech or an artificial spectral null.

Those points which were artificial nulls due to speech being removed are dealt with in section 5.3.5, all other isolated spectral peaks were then passed onto the variance classification.

5.3.3 Variance classification

The use of variance classification in this stage, was to examine all points which had passed the energy classification and determine which of these were potentially musical noise, and which could be weak unvoiced speech. In [110], the blades used covered seven spectral points in the spectrogram, but as there had been two levels of spectral classification performed before this, the variance classification could be altered.

In section 5.2.2, it was stated that without a form of energy classification before attempting variance classification, there could be small isolated sections of speech present in the spectrum. The implementation of the modified Whipple classification in the previous section, led to these points being filtered out, as they had higher spectral values than the remaining region B points. The surrounding spectral points from region A would have a larger energy value showing that there has been a loss of speech, and preventing these points from erroneously being passed to the variance stage. This left only those points which were weak energy speech, and musical noise points.

Figure 5.6 showed, that after energy classification points in region B with larger energy values were passed to the next step. The inclusion of any points other than those in the selected region, would corrupt the process, as the previous steps have made decisions based on this section of the spectrogram. As region B covered a section that was 3×3 as opposed to the 7×7 case of [110], the use of blades would not be appropriate, as they cannot be long enough to differentiate between speech and isolated musical noise points. This does not invalidate variance classification, as the points presented by region B still exhibit a variance, as described in figure 5.7.

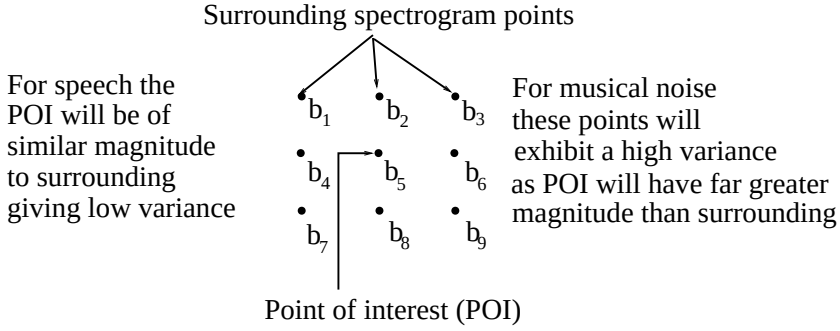


Figure 5.7: *The variance of region B*

By calculation of the region B variance, it was possible to determine musical noise points. As

the classification works in both time and frequency musical noise can be detected, given that speech points are consistent across time and frequency, meaning there was a similarity between the POI and the surrounding points. Musical noise on the other hand, is not consistent with surrounding portions of the spectrogram, due to the random placement in which the SS produces it. This gives rise to a significant variation of the amplitude in either time or frequency, allowing the musical noise point to be detected. For region B, there are nine surrounding spectrogram points, signified by $b_1, ..., b_9$ as shown in figure 5.7.

$$\mathrm{var}(RegionB_{(i,m)}) = \sum_{sp=1}^{9} \frac{20\log_{10}(b_{sp}+1)^2}{|RegionB_{(i,m)}|} - \left\{ \sum_{sp=1}^{9} \frac{20\log_{10}(b_{sp}+1)}{|RegionB_{(i,m)}|} \right\}^2 \tag{5.7}$$

where $|RegionB_{(i,m)}|$,b_{sp} are the sum of the magnitudes in the 3×3 region around the spectrogram POI and the region B spectrogram points respectively.

Equation (5.7) shows the calculation of the variance of region B which was similar to that of equation (5.4), except, that this time, it was the variance of the whole region B which was calculated, rather then variances of a number of blades which passed through the POI. Once the variance had been calculated, it was compared to a threshold as in [110], and a decision could be made as to whether the section was weak energy speech or musical noise.

5.3.4 Median calculation for musical noise replacement.

The final part of the musical noise reduction was to determine the replacement value for all of the points which had been classified as region 2 (definitely musical noise) points. In [110], the median value of the lowest variance blade was used as a replacement value for the points which were classified as region 2 points. This operation was shown in equation 5.6, where the median of the blade replaced those spectral values in which the POI was considerably larger in magnitude than the median of the blade.

Again, given that the blade approach was not being used, but rather the calculations were on region B as defined by Whipple, the median value was calculated on the nine spectrogram

points $b_1, ..., b_9$.

$$|\widehat{S}_i(m)| = \underset{(i',m') \in RegionB}{\text{median}}(|\widehat{S}_{i'}(m')|) \text{ if } \underset{(i',m') \in RegionB}{\text{median}}(|\widehat{S}_{i'}(m')|) < |\widehat{S}_i(m)| \tag{5.8}$$

where i' and m' are the time index of points in region B chosen for processing, and the frequency index of these points respectively.

As equation (5.8) shows, if a POI had a lower value than the median, it was left untouched, otherwise a replacement was chosen. The median of region B for a section of musical noise would be a considerably lower value than the musical noise point, leading to a suppression of the musical noise. If any speech points had erroneously moved through all of the classification steps, then taking the median value of region B would result in a spectral value which was related to all the surrounding speech spectrogram points, and would cause little change to the output speech POI.

5.3.5 Lost speech retrieval

In the energy classification step of section 5.3.2, it was noted that some spectral points causing region B to have lower energy than region A could be due to the removal of speech information, which would lead to spectral nulls in the spectrogram. If these spectral nulls were left uncorrected, then the resultant loss of speech information would lead to corruption of the resultant speech. This could have the effect of changing the tonal quality of the speech, if sections of low frequency information were removed from male speech, or make the speech sound muffled, if higher frequency information was removed. An example of these effects is shown in figure 5.8.

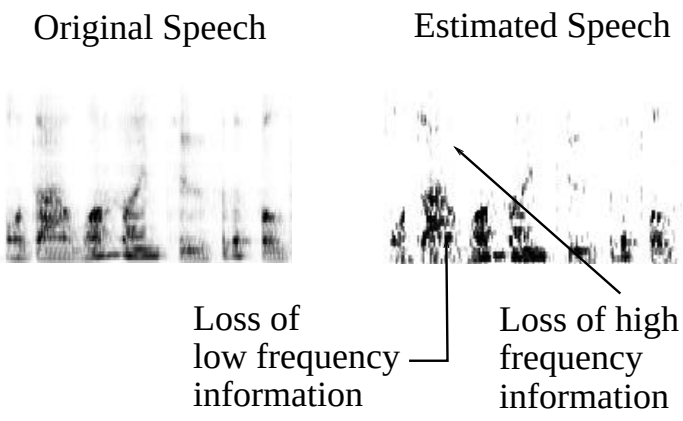


Figure 5.8: *Examples of the loss of speech information*

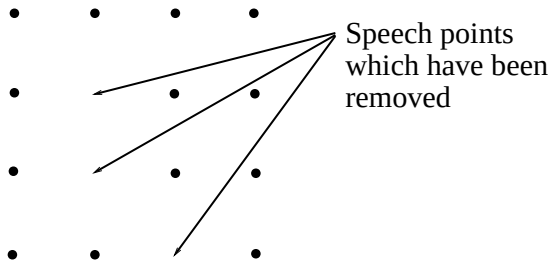


Figure 5.9: *A typical prediction case*

In order to make the speech sound as natural as possible, it was useful to be able to replace such points with spectral values, which approximate the erroneously removed speech as closely as possible. In order to do this, there needed to be some way by which the surrounding spectrogram points were examined, in order to predict the values of the points which had been removed. The standard linear predictor could be used for such a task, although it was felt that it had to be constrained to one step prediction, to prevent a succession of predictions occurring where the prediction coefficients were being calculated on previously predicted data. Doing so would cause an increasing inaccuracy of the predicted data, and lead to the insertion of spectral points, which could be as perceptually unpleasant as the musical noise which this technique tried to reduce. This meant that the one step predictor was also required, to note if current spectrogram points were the result of a previous prediction, and only attempt to update the prediction coefficients if this was not the case. A typical case where the prediction can be used is shown in figure 5.9

The standard one step forward predictor is shown in figure 5.10 where $w_1, \ldots, w_M$ and $u(n-1), \ldots, u(n-M)$ are the M tap weights and tap inputs respectively.

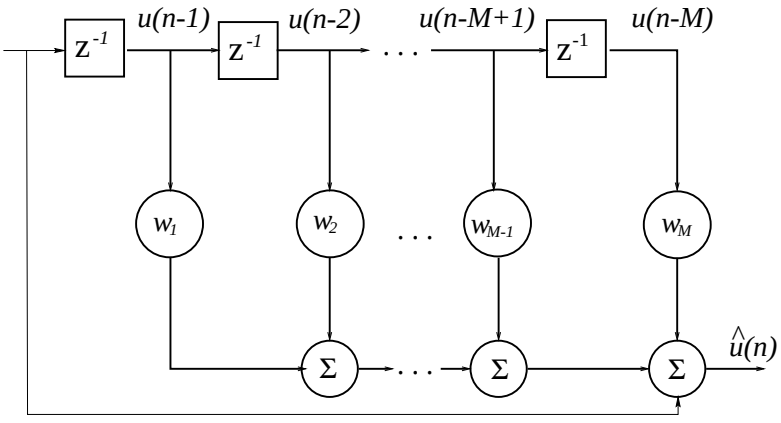


Figure 5.10: *A one step forward predictor*

From this the predicted value $\widehat{u}(n)$ can be defined as:

$$\widehat{u}(n) = \sum_{k=1}^{M} w_k u(n-k) \tag{5.9}$$

As the points being presented for prediction were spectrogram points, it could be seen in figure 5.9 that there would be spectral points of use before and after the POI. Using both of these sets of spectral points to predict the new POI would improve the accuracy of the prediction. This led to the introduction of backward prediction as shown in, 5.11 and equation (5.10).

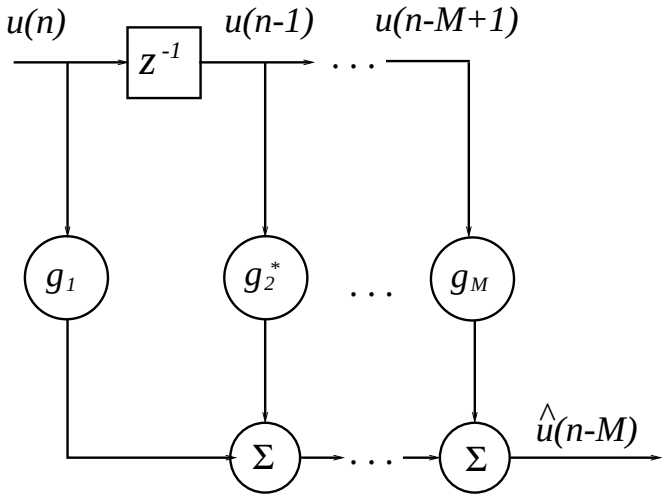


Figure 5.11: *A one step backward predictor*

$$\widehat{u}(n-M) = \sum_{k=0}^{M-1} g_k u(n-k) \tag{5.10}$$

where $g_1, \ldots, g_M$ are the weights.

For spectral estimation, the forward and backward linear predictors could be combined, to produce the *forward backward linear prediction* (FBLP), which allowed the prediction of a specific spectral point. By use of the points which surrounded the POI in either time or frequency, the FBLP can look to fill in a speech point which had been removed from the spectrogram. In order to constrain the area over which the points for the calculation of the prediction took place, it was decided to use only spectral points, which lay three points either side of the POI, to determine the furthest points which could be used in the prediction. This allowed for there to be a significant number of non-null spectral points available for prediction calculation. If only the

points corresponding to region B were used, it was likely that most of these will be zero points, making it more difficult to obtain an accurate prediction. This is shown in figure 5.12.

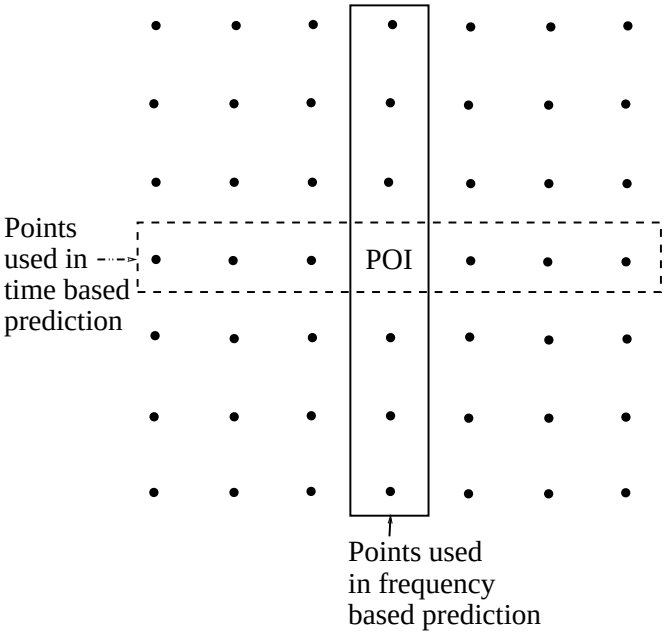


Figure 5.12: *The prediction area and some examples of the points used in the forward and backward prediction*

Before any calculations could be performed on the POI, a test was performed to determine whether the defined regions contained any spectral points which were the result of a previous prediction. Errors could occur from using a pretrained predictor, which had to use predicted data in the input data, or from the use of predicted points to calculate the prediction coefficients. In order to prevent the latter errors, the technique shifted to the use of a previous trained set of coefficients, if the input data contained previously predicted spectral points. In this case, the training phase of the coefficients was foregone.

In figure 5.12, regions for the implementation of prediction in both time and frequency are defined. In general, the method used was the time based version, although if a number of the time based points were part of a spectral null, this would result in a low (possibly zero) spectral point being placed into the spectrogram. To prevent this, the second frequency based region as defined in figure 5.12 was used instead. The technique searched to determine whether the points closest to the POI were also zeros from a spectral null, and if this condition was true, moved from using the time based regions to the frequency based. Again, this provided the technique with enough spectral points on which to base an accurate estimation. If, in the event that the technique also detected the frequency regions as being filled with nulls, then the POI

was left alone, as this portion of the spectrogram was beyond a reasonable state from which simple linear prediction could repair.

Once the POI had passed through the two classification layers, the filter weights for the forward and backward prediction could be calculated. From work in [111], it could be seen that the optimum tap weight vectors could be calculated by solution of the Wiener-Hopf equations:

$$\mathbf{Rw} = \mathbf{r} \text{ for the forward prediction} \tag{5.11}$$

$$\mathbf{Rg} = \mathbf{r} \text{ for the backward prediction} \tag{5.12}$$

where $\mathbf{R}$, $\mathbf{r}$, $\mathbf{w}$ and $\mathbf{g}$ are the correlation matrix of the tap inputs, the cross correlation matrix between the tap inputs, and the desired output and the filter weights for the forward and backward prediction respectively. As the input we were looking to predict (speech) was taken to be linear over short times, this allowed the exploitation of the Wiener-Hopf equations for the production of the tap weights. To gain the optimal tap weights the equations (5.11) and (5.12) were premultiplied by the inverse of the correlation matrix of the inputs $\mathbf{R^{-1}}$.

$$\mathbf{w} = \mathbf{R^{-1}r} \tag{5.13}$$

$$\mathbf{g} = \mathbf{R^{-1}r} \tag{5.14}$$

For the calculation of the coefficients represented in equations (5.13) and (5.14), a prediction coefficient routine from [112] which took in a vector of real data and the number of coefficients required was used. From this, the linear predictive coefficients were produced.

For the calculation of the required spectral point, the coefficients produced by the equations (5.13) and (5.14) were substituted into equations (5.9) and (5.10), along with the data points from the selected area. This produced two prediction points, one from the forward predictor, and one from the backward predictor. A linear prediction algorithm from [112] was used for both the forward and backward predictors, and the mean of these values was taken as the final

prediction point.

$$\widehat{S}_i(m) = \frac{\text{forward prediction result } + \text{ backward prediction result}}{2} \tag{5.15}$$

The taking of this mean allowed for the averaging out of any errors, which one of the predictions may have introduced, due to any spectral artefacts which could be present in the prediction data. When the replacement data point was calculated, it was then placed into the spectrogram to replace the original $\widehat{S}_i(m)$ at that point. Using this method, it was possible to replace sections of speech which the spectral subtraction algorithm may have removed, with values which would be perceptually accurate, rather than leaving spectral nulls. The software used in the linear prediction was based on algorithms provided in [112].

5.3.6 Recombination of the speech signal.

Following the application of the various sections of the musical noise reduction technique as shown in figure 5.6, the spectrogram of the entire signal could then be recombined to give the enhanced speech magnitude. Upon doing this, it was possible to introduce a degree of extra noise reduction into the technique, with little impact on the speech quality. Through examination of the spectrogram, it was possible to determine fairly accurately sections of speech/pauses, based on the degree of spectral content which each frame contained. By blanking out the frames which corresponded to pauses, this would cause the enhanced speech magnitude to stand out audibly compared to the pauses.

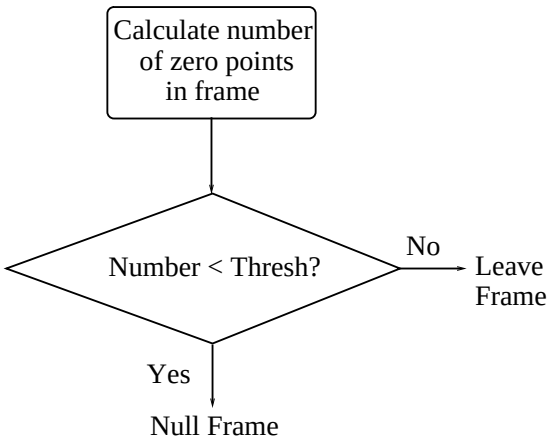


Figure 5.13: *Flow Diagram for the blanking of residual noise frames.*

The problem was trying to determine which degree of spectral content would categorise the residual background noise, as opposed to frames with low speech content from the end of utterances. The basic outline for the method is shown in figure 5.13, and by examining a frame of speech, it could be seen how the spectral content changed from frame to frame, an example is shown in figure 5.14

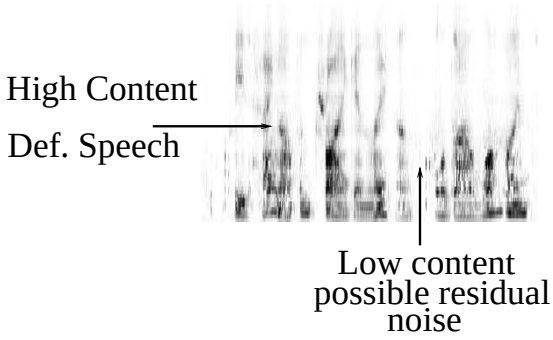


Figure 5.14: *Typical frame of speech outlining degrees of spectral content..*

As each frame holds 64 distinct frequency values (half of 128 due to Nyquist folding), this gave an upper limit of the number of spectral points which could determine whether a frame holds speech or pauses. Examples of resultant speech after selection of some arbitrary values for the threshold are shown in figures 5.15 and 5.16. More results can be seen in figures D.1 to D.4.

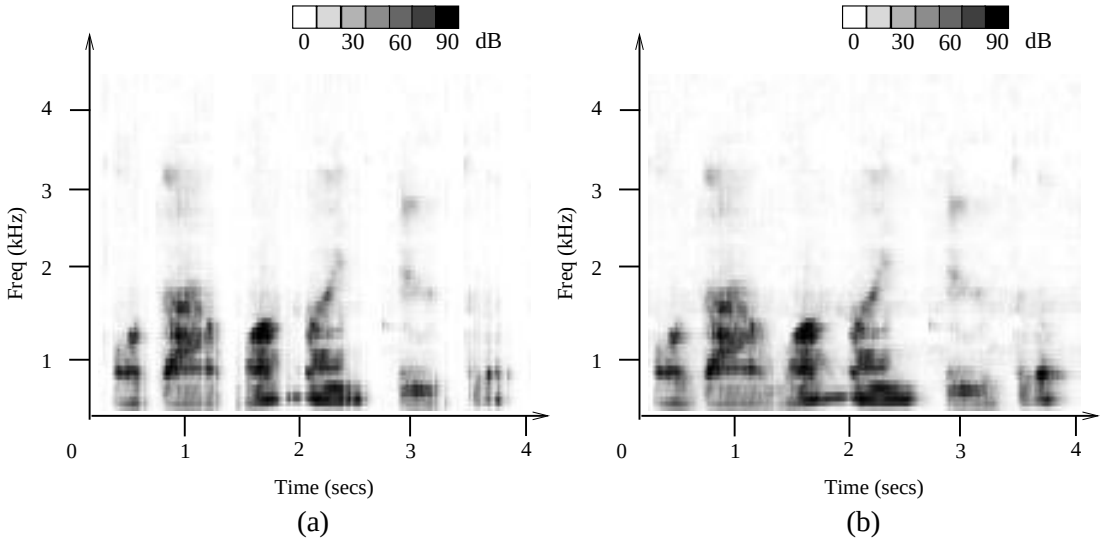


Figure 5.15: *Spectrogram of Male Speech Threshold (a) 20 (b) 50*

The test segment operated on to give the results in figure 5.15 came from a male speaker corrupted with a 1kHz tone at 0dB SNR, and it can be seen that when the threshold was set at a

low level, more frames of speech were removed as a higher percentage had over 20 zero points as opposed to 50 zero points, as in figure 5.15(b). In the female case, shown in figure 5.16, the same pattern emerges as the threshold rises. Through a set of experiments applied to enhanced speech for male and female cases over a range of SNRs, and different corrupting signals, an interesting effect was observed.

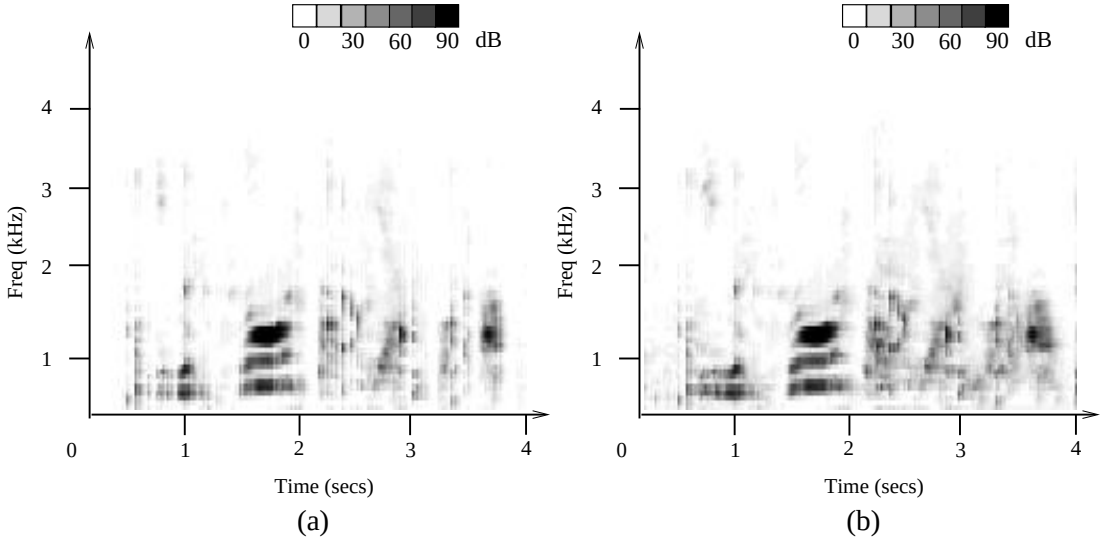


Figure 5.16: *Female Speech with a Threshold (a) 20 (b) 50*

Table 5.2 shows, the ranges of thresholds over which different test signal exhibited the best audible results for the trade off of improved background noise reduction and acceptable speech quality. The results show the point at which the best audible results were obtained, and the limit at which the loss of speech became an impairment on speech quality. It can be seen in some cases, that the range of good speech quality extended below and above the best value in terms of spectral content.

The results fell into two general categories:

1. Female speech with thresholds between 40 and 50 spectral points.

2. Male speech with thresholds between 20 and 40 spectral points.

The general trend showed that as the SNR increased, the null threshold dropped in value, but given that there was no single threshold value which could categorise the best sound quality, this led to the realisation that a different threshold value would be required for male and female speech. A choice of 40 biased the technique towards the male case, rather than providing similar

	Threshold	
Speech operated on	Best Audible	Good Audible
male/pink noise -10dB	40	20
male/pink noise -5dB	40	30
male/music 0dB	40	25
male/female 0dB	40	30
male/1khZ tone 0dB	20	40
male/female 5dB	40	25
male/music 5dB	40	20
female/male -10dB	50	40
female/male -5dB	50	40
female/music 0dB	50	40
female/male 0dB	50	40
female/1khZ tone 0dB	40	50
female/pink noise 5dB	40	50
female/male 5dB	40	50
female/1kHz tone 5dB	40	50

Table 5.2: *Null threshold ranges for differing speech conditions.*

performance for each. Unfortunately, this was a problem which there was not enough time to solve fully, and it was decided to choose values of 30 and 45 spectral points for the thresholds for male and female speech respectively. As such it was required to manually input the null threshold for each speech case, depending on the sex of the speaker. Once the nulling of non speech frames was completed, the spectrogram was reconstructed to allow the resultant speech magnitude to be transformed back into the time domain.

5.4 Listening test results

In order to determine how well the proposed musical noise reduction technique performed, it was decided to run a set of MOS tests. In order to maintain consistency, the cross section of listeners used in section 4.4.2 were again used, and the signals operated on were the same as in the listening tests of that section.

In order to judge the relative performance of the proposed musical noise reduction technique, the speech from this was presented, along with results using both the Whipple and Goh *et al.* noise reduction approaches. For both of these systems, the exact spectral subtraction conditions on which they operated on were used, so that a valid comparison could be made. Each listener

examined eight test sequences, and were given a set of instructions which outline the procedure of the test:

Listening test no 1
Removing Musical Noise from a Female Voice which has had a corrupting Male Voice removed.

The purpose of the test is to determine how well the algorithms remove the musical noise from the enhanced speech signals. Musical Noise is best described as a warbling water like sound that will be heard in the signal. On the tape you will hear 6 sections of speech. The first two give an indication of good clean speech and the noise corruption. The third section is the speech signal after enhancement and this be used as a basis for marking the tests. You do not need to do anything while listening to these.

After this, you are presented with 3 segments from different enhancement algorithms. In the table below, please enter a score for each segment as indicated on the following scale;

Rating	**Speech Quality**	**Level of Distortion**
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

The qualities you should keep in mind as you judge each segment is;

1. How good the Speech quality is.
2. How much does the distortion affect speech quality?
3. Do any of the algorithms produce a better output that the file that is presented to it (Segment 3)?

Segment Number	**R**ating
1	*Example Data Do not mark*
2	*Example Data Do not mark*
3	*Example Data Do not mark*
4	
5	
6	

If you wish to relisten to this test go to marker no 39 on the DAT.

The results of these test are shown in shown in figures 5.17 and 5.18, where again the results of the test have been split into two graphs, in order to make the results more legible. While the MOS testing of the NEAH and PANE techniques required the listener to mark the techniques on the quality of the speech and the amount of noise reduction which had taken place, in these

MOS tests the listeners had to decide how much the warbling musical noise quality had been reduced, as well as the quality of the resultant speech.

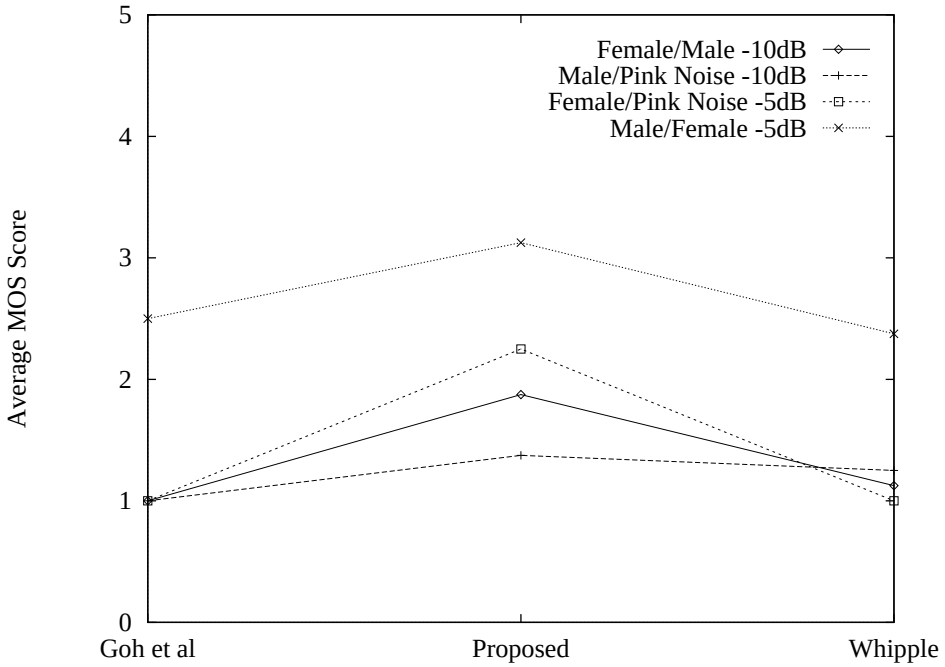


Figure 5.17: *First set of listening test results.*

Figure 5.17 presents the MOS test results for the test signals with SNRs of -10 and -5dB, which were corrupted with a competing speaker or pink noise. From these, it can be seen that of the three techniques presented, the proposed was thought to produce better quality speech and less speech distortion than the Goh *et al.* and Whipple techniques. For any kind of subjective testing, it could be expected that as the SNR of the input signal rises, there was a corresponding rise in the MOS test results, as the conditions the technique is operating on are less severe. The MOS tests are seen to rise for the proposed technique as SNR increases, although it should be noted that this was only a general increase, as the test signals being compared at each SNR were not the same. In comparison, the Goh *et al.* and Whipple provided almost equal performance over the range of SNR and signals, which was a desired characteristic of any technique, although unfortunately the score could be seen to be consistently low, indicating poor performance. The closest performance either of these techniques attained, compared to the proposed case, was in the Male/Pink Noise scenario. Overall, it could be seen that the proposed technique improved from just above *unsatisfactory* to *fair* performance for an increase of only 5dB, which given that the test condition was very severe, was a good measure of the performance. The Goh *et al.* was felt to perform equally across these test conditions, with an average MOS

score of 1.

It was interesting to note, that in these tests for the proposed musical noise reduction technique that the female speaker was seen to offer better results in pink noise than the male, and the opposite seemed to be true for the competing speaker case. Comparing this to the Whipple technique, it could be seen that for this set of test results, the Whipple technique tended to favour male speech with both tests resulting in higher average MOS scores for male speech. The results for the proposed technique are encouraging, as they showed that there was no overall bias in the performance towards male or female speech enhancement, which was important as an enhancement system which could work for only one sex will be of little use.

Results for conditions which were more likely to be experienced by a mobile phone user in terms of the SNR were presented in figure 5.18. In these test cases, the corrupting signals were either the 1kHz tone or the section of music. The most noticeable aspect of these results was the noticeable shift up the MOS ratings for the average results of each case.

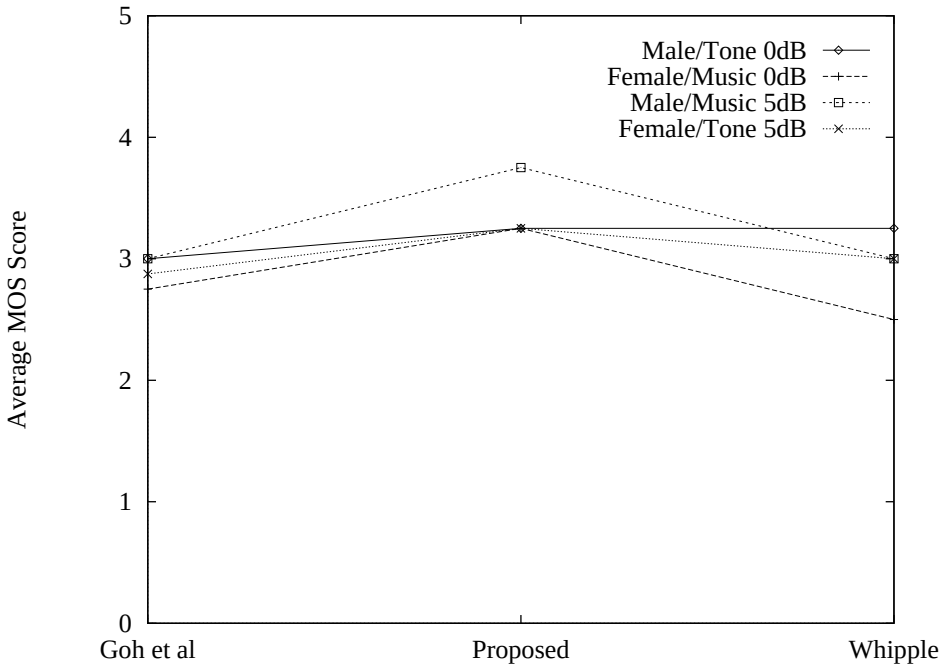


Figure 5.18: *Second set of listening test results.*

It can be seen, that for the proposed technique, the listeners rated three of the four test cases as producing *fair* quality speech with the fourth being rated closer to *good.* Once again, it can be seen that overall the proposed technique tends to operate equally for male and female speech, with no particular bias. The Goh *et al.* approach in these test cases seemed to result in slighter

higher average scores for male speech, similar to the Whipple technique. Of interest was the results obtained for the male/tone 0dB SNR case, where the Whipple technique produced an output that was deemed to outperform the proposed technique. It was felt that this could be due to the applied value of the null threshold value, which for the male speech was taken to be 30 zero spectral points across the range of conditions. In 5.2, this was the only case where the best audible results could be achieved, with a low threshold value, and setting higher than this level resulted in a corresponding drop in performance. The Whipple technique on the other hand did not include the nulling technique, and the simple energy method of [109] performed better.

The overall impression from the MOS tests results was that the proposed musical noise reduction technique outperformed the constituent parts which made up the basis of the technique. The listeners consistently felt that the proposed technique provided the best quality speech, with as little of the warbling effect of the musical noise present as possible.

5.5 Summary

In this Chapter the effects of musical noise on enhanced speech were examined. It was shown that musical noise could result in severe distortion of processed speech, which made it difficult to understand and unpleasant to listen to. Some classification techniques to recognise musical noise were discussed, and the work of G Whipple and Goh *et al.* examined, which used the methods of energy and variance classification. It was seen that by using either of these, sections of musical noise could be detected and removed, although neither provided a complete solution to the problem. In order to find a solution, a new musical noise reduction technique was presented, which pieced together aspects of the two classification methods by restricting the work of [109] to a smaller region, and using this area for some variance classification motivated by the technique in [110]. By applying the technique in [109] for the original 7×7, and a reduced 5×5 area, it could be seen that the errors produced by the 5×5 were smaller than those of the original 7×7 case in [109]. This showed that the use of a smaller spectrogram region was justified.

After the energy classification was carried out, it was shown that this restricted area prevented the use of a blade based approach, as the blades would not be long enough to differentiate between speech and isolated musical noise points. Despite this, the underlying variance classification of [110] could be used, to determine whether a POI was speech or musical noise. It

was shown through examination of the surrounding spectrogram points that an isolated musical noise section could be detected. A median calculation was performed on the spectrogram area used for the variance classification to replace musical noise sections. This reduced large amplitude values to a level closer to the surrounding spectrogram, making the speech sound more natural.

It was shown that SS could result in areas of spectral nulling in regions of the spectrogram. This reduced the quality of the processed speech, by removing speech information, and this had to be redressed. In order to replace these nulls with a perceptually relevant substitute, it was decided to use a linear predictor which could be trained on the surrounding spectrogram points in either time or frequency. A FBLP was constructed from techniques in [112] and the training process was automatically supervised to prevent training on previously predicted data, which could compound errors in the substituted value.

As a final attempt to make the resultant enhanced speech stand out audibly compared to the speech pauses, those frames which contained only noise were nulled. A threshold was defined which determined how much spectral content was classified as residual noise only, and how much was actually speech. Surprisingly, it was found that there was a discrepancy between the average threshold required for male and female speech, and unfortunately there was not time to implement an automated null routine, which could determine between the apparently differing needs of male and female based enhanced speech. Manual inputting of the threshold value was required for each test signal, in order to use this technique.

A set of MOS listening tests were performed for the same test signals used in section 4.4. For consistency, the same panel of listeners were used, and the tests were randomised. The results of these tests showed that the developed musical noise reduction technique performed fairly well over the range of test signals, outperforming both the work of [109] and [110], on which some of the work was based and was able to replace sections of speech which had removed by the SS process. The overall MOS tests were are not extremely high, due to the low SNR of the signals being operated on, although it was hoped that when this system was integrated into the overall enhancement algorithm, the real benefits of the technique would be seen.

The integration of this process into the overall speech enhancement algorithm results in the process shown in figure 5.19

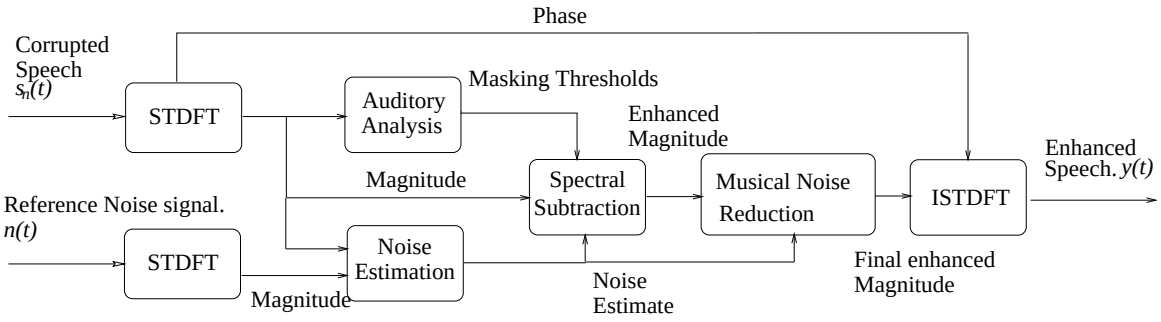


Figure 5.19: *Flow Diagram of how the musical noise reduction criterion are integrated into the speech enhancement.*

Chapter 6
The Estimation of Clean Speech Masking Thresholds.

In this Chapter, the use of masking thresholds for clean speech is further examined. The use of the psychoacoustic criterion of Chapter 3 requires an accurate estimate of the clean speech masking thresholds, in order to determine which portions of noise are masked, and which are not. In this Chapter, the typical technique for the estimation of clean speech masking thresholds is examined, and it is seen that the estimation can be affected by the presence of musical noise or spectral nulls. This can result in masking thresholds which are not accurate, causing perceptual based speech enhancement to operate poorly. A new method of estimating the clean speech masking thresholds is examined, which does not operate on an estimate of the clean speech, but examines masking thresholds produced by the corrupted speech and the corrupting noise. This Chapter is an extension of work which the author presented in [113].

6.1 Why do clean speech masking thresholds need to be estimated?

The advantages of applying auditory masking thresholds to any form of audio signal processing is obvious. In applications such as the Mini-Disc, thresholds allow the codec to determine which portions of a musical signal are audible, and which are not. Those tagged as being inaudible (under the threshold) do not require to be coded, and through application of coding techniques, this is where part of the data compression can be achieved. In speech enhancement masking thresholds can allow an increase in algorithm speed, by operating only on audible points of the signal. A saving will only be achieved, if the computation required for calculating the thresholds is less than was required for SS of these points in the first place. If this is not the case, the small computational increase can be accepted, when the result is improved speech quality, as the algorithm now takes account the way the ear hears. In noise shaping, masking thresholds manipulate any residual noise into a more perceptually acceptable form.

In speech enhancement, perceptual techniques encounter problems, as there is normally no

clean speech reference from which the thresholds can be calculated. To compound this, there is the need for a good spectral estimate of the noise, to ensure that the performance of the algorithm is maximised. The better the estimate, the lower the chance of speech distortion from the spectral subtraction. In a stationary noise environment this can be readily done, but the more non-stationary the noise, the harder this becomes.

For perceptually based spectral subtraction algorithms, one of the most important aspects is an accurate estimate of the clean speech masking threshold. As can be seen in figure 6.1, if the estimate is inaccurate, the enhanced speech may either contain a high degree of residual noise, or cause a loss of speech information.

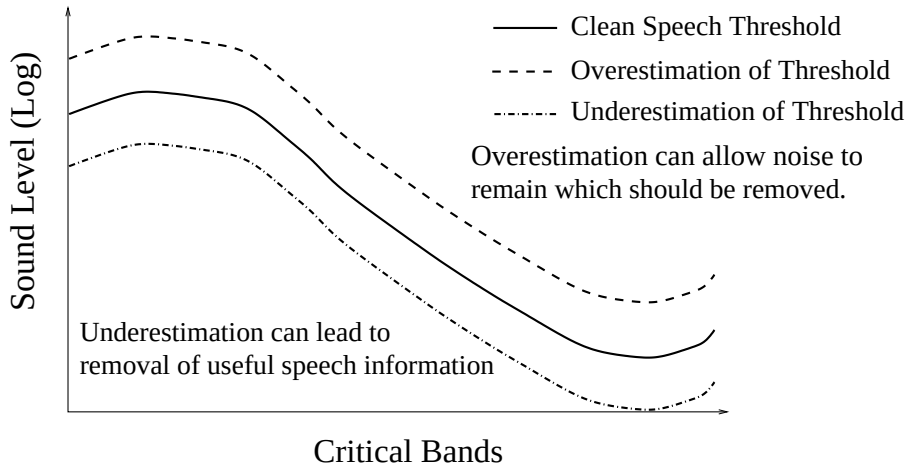


Figure 6.1: *Example of the problems of Clean Speech Threshold Estimation.*

Underestimation of the masking threshold leads to the false premise, that more sections of the presented signal can be heard than is actually the case. As the masking thresholds determine which portions of the signal are subject to the SS, this could lead to sections of speech being incorrectly assumed to be audible, and SS applied to them. While the original speech magnitude may not have been audible, there is the chance that the musical noise which can arise from the enhancement shall be audible. This could lead to a reduction in the speech quality, and removal of speech information. This also leads to an increase in the computation time required for the SS, as more spectral points are deemed to be audible.

In the case where the clean speech masking threshold (CSMT) is overestimated, the converse is true. The estimated CSMT is set too high, and points which were audible and candidates for the SS are now deemed to be inaudible, and not passed to the SS. While this can lead to an increase in the performance of the SS in terms of the computation time (as less spectral

points pass through the SS), it has the disadvantage of resulting in more noise remaining in the resultant speech, which should have been removed. This causes a drop in the noise reduction performance of the whole enhancement algorithm, and in an extreme case, could lead to the speech being no more audible than the original corrupted case. The ideal estimation would of course result in an exact mapping of the estimated CSMT onto the actual CSMT, but in practise, it is more likely to be in between these two extremes, with the shape and level of the actual CSMT mapped as accurately as possible.

6.2 Calculation of the CSMT

In section 3.3, the description of the integration of perceptual criterion into the SS speech enhancement algorithm was given. In this, it was shown that standard masking threshold calculation required an estimate of the enhanced clean speech. In any noise reduction environment, there will be no access to the clean speech signal, meaning that an estimate of the CSMT must be made. The process is shown in figure 6.2

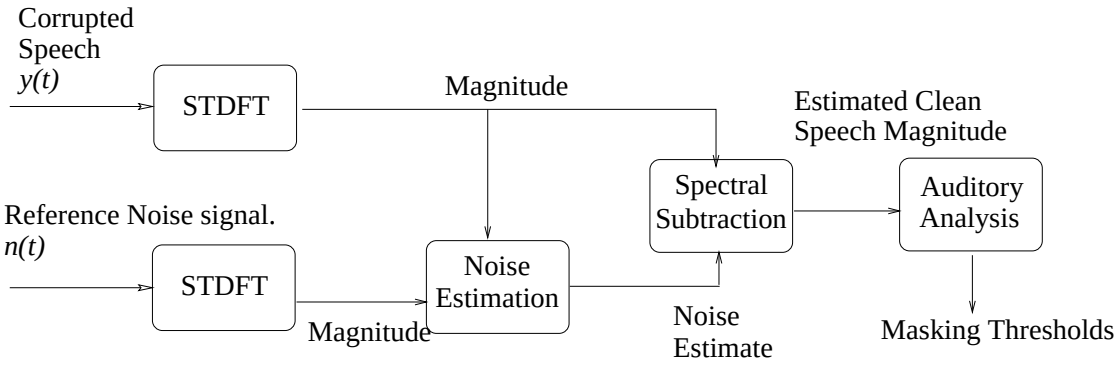


Figure 6.2: *Example of the typical estimation of CSMT.*

In figure 6.2 it can be seen that SS is normally performed on the corrupted speech, to gain an estimate of the clean speech from which the estimated CSMT can be calculated. A problem arises if the clean speech estimate is not accurate, then the masking thresholds which are derived from it can never be accurate. As seen in Chapters 4 and 5, even with a good noise estimate, it was still possible to obtain elements in the enhanced clean speech which have no relation to the original clean speech. The more adverse the noise conditions and SS could lead to the production of musical noise artefacts, as shown in Chapter 5. These can affect the calculations of the estimated CSMT in two ways:

1. The random removal nature of spectral subtraction can result in large energy peaks in the magnitude spectrum (as described in section 5.2). As the magnitude spectrum is used to calculate the masking threshold, this has the result of artificially raising the critical band energy (and hence masking threshold), associated with the affected frequency range. This results in the masking thresholds being set too high, and during enhancement noise that should be removed, is erroneously tagged as being inaudible as shown in figure 6.1

2. If a number of points in the magnitude spectrum are found to be below zero after SS, the use of half wave rectification will lead to these being assigned a zero spectral value, resulting in spectral nulls. The same effect although not as pronounced occurs if these negative SS values are replaced by a noise floor. If a critical band has a number of spectral points which are set to nulls, this reduces the CB energy and has the effect of reducing the masking threshold (or in the case of half wave rectification producing errors), which can cause the removal of useful speech information as shown in figure 6.1.

The task was to either obtain a more accurate $\widehat{S}_i(m)$, or not use this as a method of calculating the estimated CSMT. Having spent a great deal of time examining methods of speech enhancement it was obvious that the quest for a more accurate $\widehat{S}_i(m)$ was not the optimum path to take. At the end of the day, unless the estimate was near perfect, errors would still be passed onto the masking threshold calculation. If an ideal estimate of the clean speech could be produced, this would not require any of the extra processing, and would be of great use. It was decided that a new method was required which could sidestep the corruption of the SS produced thresholds.

The only signals which were presented to the speech enhancement algorithm, were the corrupted speech magnitude $Y(m)$ and the background noise magnitude $N(m)$. It had already been seen that trying to examine the relationship between these magnitudes to determine $\widehat{S}(m)$ could be problematic. On the other hand each of these signals, if taken in isolation, would produce a set of masking thresholds corresponding to the excitation the signal imparts. If a set of masking thresholds were produced for the corrupted speech, and the estimated background noise $\widehat{N}(m)$, then by examination of the relationship between these, it may be possible to produce an estimation of the CSMT, which did not suffer from the corruption that could occur from SS. The proposed solution is outlined in figure 6.3.

It can be seen in figure 6.3 that the masking thresholds of both the corrupted speech and the estimated noise were calculated. This was done by the perceptual routine described in section

3.3. These thresholds were then compared to see the relationship between them. From this, a weighting coefficient was produced for each critical band in the frame, and this controlled energy subtraction of the two thresholds. At all stages, the spectral points were being manipulated in the non-linear auditory based critical band domain. A three step median filter is used, to smooth out any frequency domain transients that may occur due to the subtraction process.

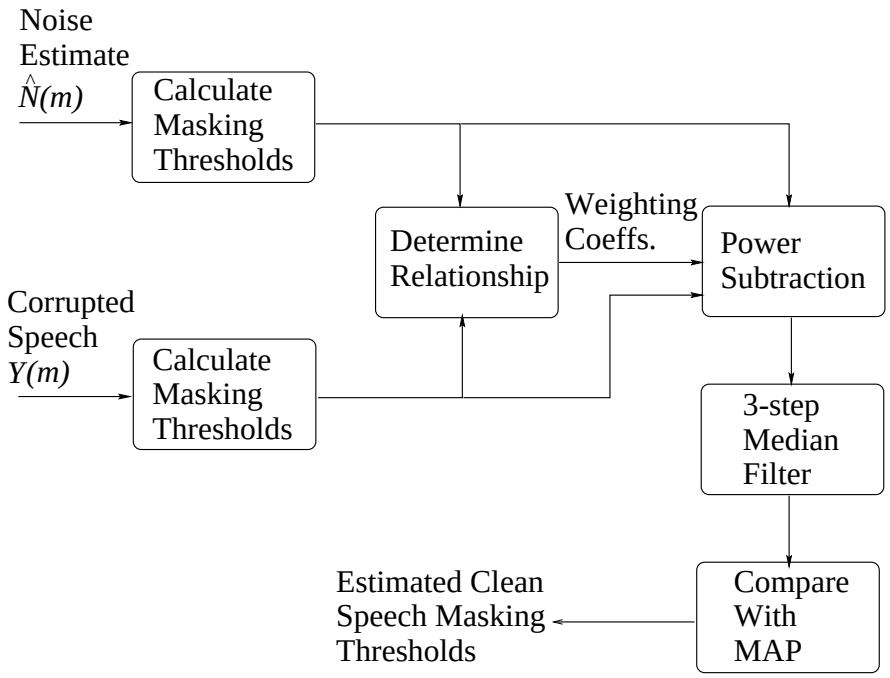


Figure 6.3: *Block Diagram of the proposed method for estimating CSMT.*

6.2.1 Calculation of the weighting factor

The most important aspect of figure 6.3 was the examination of the relationships between the thresholds produced by the corrupted speech, and the estimated background noise. For each critical band, the excitations of $Y(m)$ and $\widehat{N}(m)$ caused different threshold values to be calculated. Between these thresholds it was possible to determine an estimate of the CSMT given that $Y(m)$ held the thresholds of speech plus noise, and $\widehat{N}(m)$ held only the information for the estimated background noise. There was an energy difference in the critical bands where speech and noise were present, compared to an estimate of only the estimated background noise. An example of this can be seen in figure 6.4, which shows the interaction of the $Y(m)$ and $\widehat{N}(m)$ thresholds for a typical frame of a signal, which a mobile phone user may produce. As with all the threshold diagrams shown here, the frame used was taken from the midpoint of the test

signals, where the interference of speaker and background noise was most pronounced.

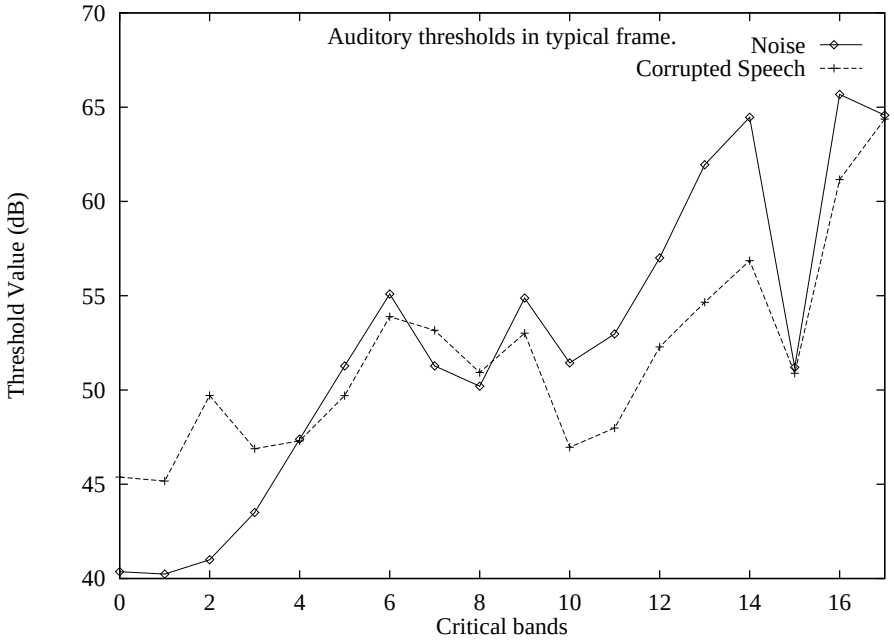


Figure 6.4: *Example of the relationship between the masking thresholds of* $Y(m)$ *and* $\widehat{N}(m)$.

It can be clearly seen in figure 6.4, that the interaction between the two sets of masking thresholds for each CB falls into the general categories:

- $\mathrm{SNR} = 0\mathrm{dB}$
- $\mathrm{SNR} < 0\mathrm{dB}$
- $\mathrm{SNR} > 0\mathrm{dB}$

A simple energy subtraction would not always work, as points where the thresholds were equal would end up with null thresholds (obviously not the case), whereas cases where one threshold had more than twice the energy of another would hardly cause any change in the final value. An technique was required, which could determine how the two thresholds changed from CB to CB, and applied a weighting depending on the condition being observed. This suggested that the general areas mentioned above required different weighting factors, to gain adequate estimation of the CSMT.

Is was difficult to decide whether better results could be obtained, if the threshold estimate contained sections which were lower or higher than the original CSMT. As previously stated,

these conditions had an impact on the quality of the enhanced speech, and the degree of noise reduction which the technique could achieve. It would be naive to assume that any estimation technique would produce a perfect CSMT, especially given the non-stationary background environment which the technique was performing in. It was deemed that adequate estimation of the thresholds would consists of a close tracking of the original CSMT, with no particular bias above or below the original.

Upon experimentation with the general SNR categories previously stated, it was found that these were not adequate enough to cover the possible interaction that each CB could experience. The points around 0dB required considerably different subtraction conditions from those with a SNR > 3 dB or < -3 dB. These points were the most sensitive to large changes in the estimation by the energy subtraction, as can be seen in figure 6.5.

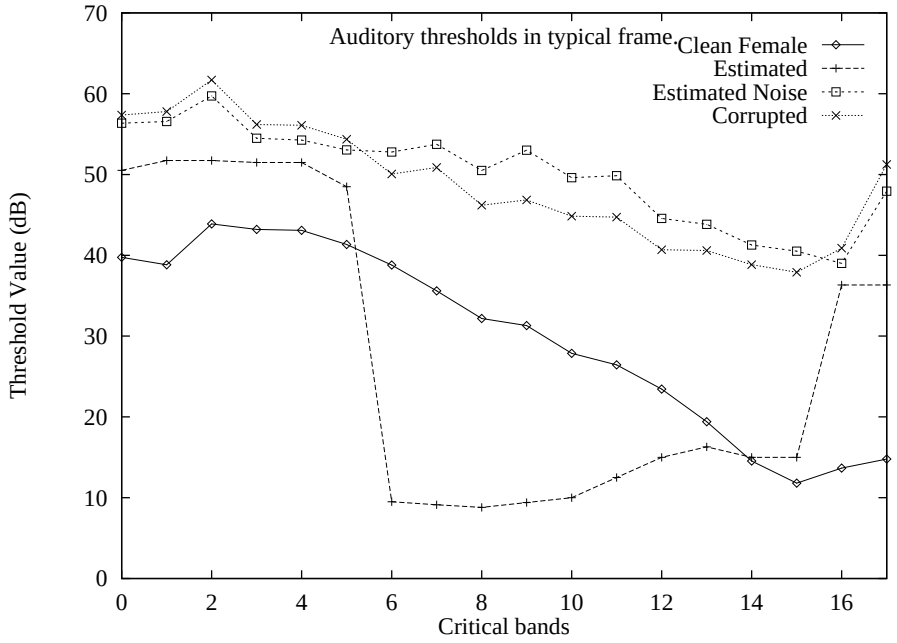


Figure 6.5: *Example of how the estimation process can change for CBs with a small SNR.*

Figure 6.5 shows the masking thresholds as calculated from $Y(m)$, $\widehat{N}(m)$ along with that from the energy subtraction if a straight energy subtraction, were applied as described by:

$$\widehat{E}_{\widehat{S}}(m) = \begin{cases} E_Y(m) - \delta E_{\widehat{N}}(m) & : \quad E_Y(m) - \delta E_{\widehat{N}}(m) > MAP \\ MAP & : \quad \text{otherwise} \end{cases}$$

where $\widehat{E}_{\widehat{S}}(m)$, $E_Y(m)$, $E_{\widehat{N}}(m)$, δ and MAP are the estimated clean speech energy, the corrupted speech energy, noise energy, weighting factor, and minimal audible pressure (absolute threshold) respectively. After conversion back into decibels, this gave the estimated CSMT. It can be seen in figure 6.5 that applying $\delta = 1$ for all CB resulted in an estimation which varied from CB to CB. It was particularly interesting to see how the estimated CSMT changed rapidly when the threshold of $\widehat{N}(m)$ moved above that of $Y(m)$ for even a small SNR. This led to a more specific categorisation of the interactions which could occur between the thresholds:

1. $\mathrm{SNR} = 0\mathrm{dB}$
2. $|\mathrm{SNR}| \leq 1\mathrm{dB}$
3. $1 < |\mathrm{SNR}| \leq 3\mathrm{dB}$
4. $|\mathrm{SNR}| > 3\mathrm{dB}$

Case 1: $\mathrm{SNR} = 0$ **dB**
In this condition $Y(m)$ and $\widehat{N}(m)$ were deemed to be equal in value, a condition which was most likely to occur in those frames and CBs where there was no speech present, and $Y(m)$ would only consist of the background noise. In this condition, the $\delta = 1$ condition could be applied, causing full energy subtraction to be applied in the subtraction equation.

As with all the above categories to counter the effects of threshold nulling, the calculated thresholds were compared with the absolute threshold for that CB, as defined in section 3.3.5 and [102] as shown in the subtraction equation. If any calculated threshold values were less than the absolute threshold, they were replaced. This would work particularly well in frames where only noise was present, as it presented the maximum number of spectral points to the SS for enhancement during non-speech activity and maximised the opportunity for noise reduction in these frames.

Case 2: $|\mathrm{SNR}| \leq 1$ **dB**
It could be seen from figure 6.5 that the CB thresholds which lay in this range (which did not include the $\mathrm{SNR} = 0$ dB case CBs) suggested that there was a significant speech threshold portion in these CBs (as the threshold $\mathrm{SNR} \neq 0$ dB), which could not be left unenhanced. It could be seen though that applying $\delta = 1$ could lead to the estimation process collapsing and the MAP values being substituted. This would not be the ideal case, as the contribution of the

speech in these CBs would not be accounted for, leading to a masking threshold that was too low as in CBs $6-13$ in figure 6.5.

On the other hand, it could be seen in CBs $1-6$ that if the subtraction equation were not weighted adequately enough, then the threshold subtraction would not remove enough of the influence of the noise, leading to an estimated CSMT, which would be too high. In both of these cases, it was important to ensure that the influence of the speech masking threshold was emphasised, as it was this threshold which was of interest. Oversubtraction of the noise threshold would not be a viable option, as this would lead to possible nulling of the calculated threshold, and result in a final threshold which had the majority of values replaced by MAP values.

It was found through experimental alteration of δ that the CB threshold interactions which lay in between $0 < \mathrm{SNR} \leq 1$dB were affected in a similar manner to the thresholds points which lay in between $-1 \leq \mathrm{SNR} < 0$dB. By applying the same δ to these threshold points, it would simplify the categorisation process. The δ value which allowed for the closest general tracking of the clean speech thresholds was $\delta = 0.8$.

Case 3: $1 < |\mathrm{SNR}| \leq 3$ **dB**

While examining the appropriate weighting factor for the $-1 \leq \mathrm{SNR} \leq 1$ dB, it was noticed that this weighting had little effect on points which lay outside this range. Figures 6.4 and 6.5 showed that there may be a larger number of threshold points which lie in the range $1 < |\mathrm{SNR}| \leq 3$ dB than $-1 \leq \mathrm{SNR} \leq 1$ dB. It was also true that these points tended to cluster together, such as from CB $8-18$ in figure 6.4 or CB $6-15$ in 6.5. When the outer limit of this category was reached, one of the critical band thresholds had twice the signal power of the other, and it would not be possible to set a hard δ decision process for such a wide range of points given the potential swing in power.

To overcome this, it was decided to allocate δ in this category on a soft decision basis, allowing the estimation to adapt to the changes in relative threshold energy. The difference between the corrupted and estimated noise thresholds was calculated, and this told which threshold held the most energy:

$$diff_{th} = Th_{Y(m)} - Th_{\widehat{N}(m)}\mathrm{dB} \tag{6.1}$$

where $diff_{th}$, $Th_{Y(m)}$, $Th_{\widehat{N}(m)}$ are the difference between the thresholds, the corrupted speech CB threshold, and the estimated noise CB threshold respectively in decibels. From 6.1, it was possible to tell whether the SNR of the CB was positive or negative. This was important as a difference of 3 dB meant that the corrupted speech threshold had twice as much power, whereas a difference of -3 dB meant the estimated noise threshold had the greater power. Using this information, the energy ratio of the critical bands could be normalised over the whole CB, taking 2(3dB) or 0.5($-$3dB) to be the maximum, depending on the calculated difference.

$$E_{nor}(m) = \frac{E_{ratio}(m)}{\zeta} \tag{6.2}$$

where $E_{nor}(m)$, $E_{ratio}(m)$, ζ are the normalised energy ratio, the energy ratio between the corrupted speech thresholds, and the estimated noise thresholds, and the maximum limit (either 0.5 or 2) respectively.

This gave a value of $E_{nor}(m)$ between $0 \rightarrow 1$ depending on the threshold energy ratio. If this value were simply applied to the subtraction as the weighting factor, then nothing more would have been achieved than the alteration of the signals to redress the energy mismatch between the thresholds, leading to a threshold null which would again be replaced by the relevant MAP value. This obviously would be of little value in obtaining an accurate estimated CSMT, although some consideration must be made of the relative changes in energy values between the corrupted speech and estimated noise thresholds. In order to do this, it was again decided to emphasise the contribution of the corrupted speech thresholds, by reducing the weighting of the estimated noise thresholds by half:

$$\delta = 0.5(E_{nor}(m)) \tag{6.3}$$

Case 4: $|\mathrm{SNR}| > 3$ **dB**

The only CB threshold sections which fell outside any of the other classification regions were those where there was a considerable difference in the threshold values, due to the influence of speech or noise present in the CB. If the SNR was positive in this section, then it could be said with confidence that there was a high degree of influence by the speech information in the CB.

Conversely, a negative SNR showed that there was very little speech, and that the estimate of the noise was not as statistically precise as hoped. When the threshold ratio reached this level, it became difficult to reduce the contribution of the noise energy from the estimated CSMT, although it was still possible to follow the general pattern of the original CSMT. The advantage of being able to do this, was that while it may be difficult to determine the baseline level of the thresholds, if the weighting effect was kept constant over all CBs, then perceptually the effect was consistent.

This effect is demonstrated in figure 6.6, where two estimations of the CSMT are shown. One of these has the same general shape as the CSMT, but is at an incorrect level. The second does not have the general shape of the CSMT, as well as being at the wrong level. This shows that of the two conditions, it was preferable to have a masking threshold which held the correct shape, even if its level was incorrect, as this prevented the auditory based speech enhancement from artificially emphasising one portion of the spectrum in a manner that was not consistent with the ear. A threshold which had the correct general shape but wrong level would result in a degradation, which would be perceptually consistent for the ear meaning the speech would continue to sound natural, if slightly noisy.

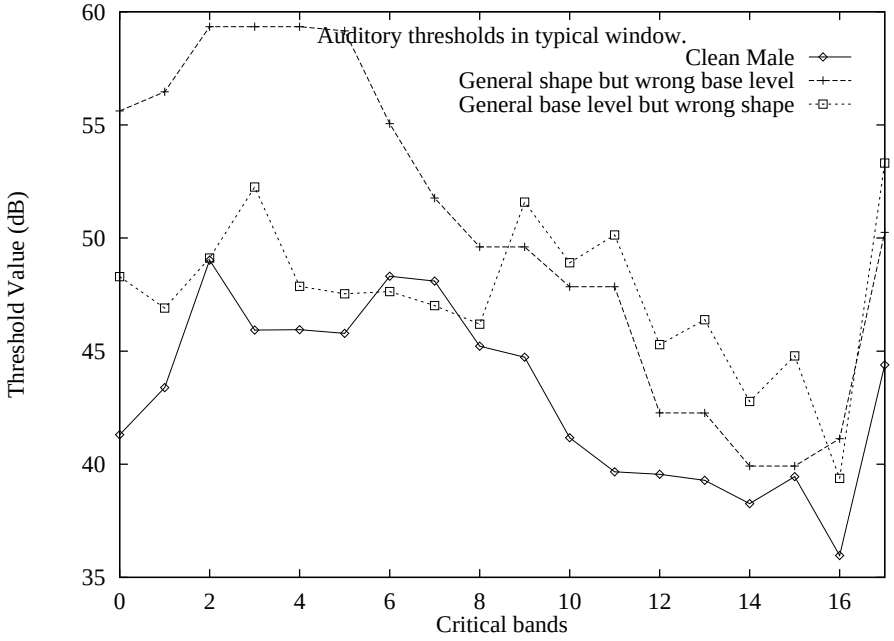


Figure 6.6: *Example of the differences between inaccurate threshold base level and threshold shape.*

The method of CSMT estimation in this case was therefore taken to only include the dropping in

level of the corrupted speech threshold, to remove some of the energy effects of the background noise. A factor ϑ controlled the drop in energy level of the threshold which was set at a nominal value of $\vartheta = 4$ for $\mathrm{SNR} > 3\mathrm{dB}$ and for $\mathrm{SNR} < -3\mathrm{dB}$, $\vartheta = 3$.

6.3 Performance of the CSMT estimator

To determine how close the CSMT estimates were, the technique was applied to a number of test signals with various corrupting speakers and SNRs. A graphical method was deemed the most appropriate way of examining how well the technique estimated the masking thresholds, for the reasons outlined about the difference between obtaining the correct masking shape and the correct base masking level. A plot was taken from a typical frame lying exactly in the middle of the test pattern which displayed the original speech thresholds for that window along with those from the CSMT estimation, and those calculated by the more common method of estimating the clean speech $\widehat{S}(m)$, and obtaining the masking thresholds from that.

A selection of the results are presented here, to illustrate the technique performance. A more comprehensive set can be found in Appendix F, which covers all of the test signals used, with particular emphasis placed on conditions where the SNR was $0\mathrm{dB}$ and above. These were more practical test conditions than the very low negative SNR cases. This does not mean that those results are invalid though, as they demonstrate how such an technique would perform in such a low SNR non-stationary environment. Figures 6.7 (a) to (d) show a sample range of the performance CSMT estimation.

It should be stressed, that these results represent only one window from a possible selection in the entire signal but they give an indication of the capabilities of the technique. It can be seen in figure 6.7 that the estimation accuracy did vary according to the type of corruption and corruption level experienced. In figure 6.7, some results for the CSMT estimation are presented along with the thresholds for the power SS approach, and the original clean speech. In figure 6.7 (a), it can be seen that the CSMT estimation produced a set of thresholds which tracked the clean female masking thresholds up to CB 13, where they diverged. The power SS results in figure 6.7 (a) produced a set of thresholds which did not mirror the actual clean female value until CB 14.

The divergence of the estimated thresholds at high CBs meant a possible increase in the perceived background noise at higher frequencies. On the other hand, the SS produced would

result in a greater cross section of the speech being presented to the enhancement algorithm. This would lead to a higher degree of musical noise and possible loss of speech information. The performance of the estimation technique was encouraging, given that the corrupting signal for this case (a male speaker) was 10dB higher than the desired female speech.

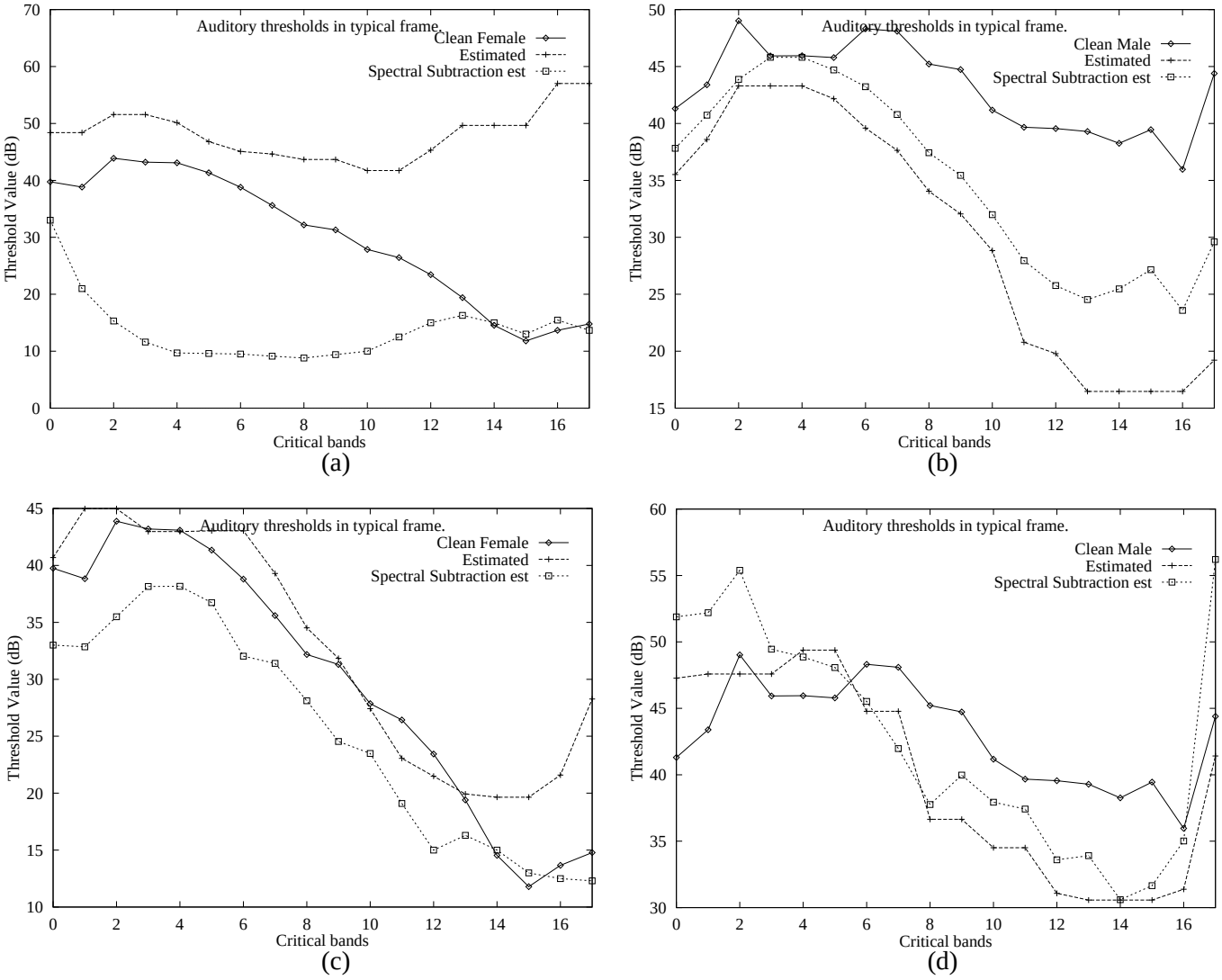


Figure 6.7: *CSMT estimation for (a) Female in 1kHz tone corruption SNR 10dB (b) Male in Female speech corruption SNR −5dB (c) Female in male speech corruption SNR 0dB (b) Male in pink noise corruption SNR 5dB*

Figure 6.7(c) shows the results of the estimation, in the case where the corrupting signal was a male speaker and the SNR is 0dB. Here it could be seen that the CSMT estimation technique provided a closer match to the actual thresholds, than the power SS case. It was slightly more conservative in this test case, as the estimation allowed more potentially noisy points to be left

unaffected by the SS. This was acceptable, as a small amount of noise in any resultant enhanced speech was thought to reduce the possibility of musical noise as discussed in section 5.2. Again there was some divergence from the actual thresholds at the higher CBs, but overall in this case, the proposed CSMT estimator was able to produce an acceptable estimation of the female thresholds.

The results in the cases where male speech thresholds were calculated, show a different trend emerging. Figure 6.7(b) shows the results of the estimation, and power SS thresholds for a test case with a female speaker at a SNR -5dB. For both the power SS and the proposed CSMT estimation technique, it can be seen that they had difficulty estimating the actual male threshold. The estimation results produced by both were similar in both level and shape, with the power SS approach giving the better result in terms of threshold shape. A similar condition can be seen in figure 6.7(d), where the proposed CSMT technique produced an estimate which was more accurate up to CB 8, at which point the power SS produced a slightly better estimate. There was not much difference between the performance of the estimation techniques, although the proposed CSMT estimator did produce a more conservative estimate again for the lower CBs.

Looking at the results in figures E.1 to E.6 in Appendix E, it can be seen that there was a definite difference between the performance of the CSMT technique, with regard to male or female threshold estimation. An example of this can be seen in figures E.3(a) and E.4(a), which show estimation results obtained from the same test signal, with figure E.3(a) showing the female estimation, and figure E.4(a) showing the male estimation. In the cases where the estimation was performed on a female voice, the proposed CSMT technique generally performed equally well or better than the normal power SS threshold method. This held across the range of signals used as background noise, as well as through the range of SNR used. As the SNR increased, there was a improvement of the estimation results for both the male and female thresholds, but the basic difference between the accuracy of the estimation between the sexes held. The cases where the best estimation of male thresholds were observed can be found in figure E.1 and E.3(c). The results gained for figure E.1 were surprising, in that the SNR was very low (-10dB), and the estimation performance for the female speech was not as accurate.

The best performance measure of the proposed technique due to the sex of the speaker, was shown by comparing all the test cases where the proposed CSMT estimator produced an accurate output. The estimate was deemed to be accurate, if the performance was equal to or better than the power SS case. In all, there were 24 test cases produced, 12 for each sex of speaker.

The results showed that:

1. For female speech, 8 test cases produced an accurate threshold estimation.
2. For male speech, 3 test cases produced an accurate threshold estimation.

This result showed that the CSMT estimation technique performed far better in estimation of female masking thresholds, as opposed to male thresholds. This was an interesting result, as there seemed to be no basis for the increased performance of female threshold estimation, over equivalent male cases. The techniques applied to all the test cases were identical, (application of DFT, then the CSMT technique), yet there were visible differences in the performance of the CSMT technique. It can be clearly seen from these results, that the female estimation was superior to the male results, which mirrored the results in section 5.3 of there being a difference in the performance of the technique based on the sex of the speaker.

6.4 The discrepancy between male and female threshold estimates.

In order to determine whether the results produced by the CSMT estimator were valid, an investigation was made of the perception of speech by humans. If it could be theorised that human hearing could perceive one sex better than the other in additive noise, this would explain the apparent discrepancy between the male and female results from the CSMT estimation technique.

The first step was the classification of which speech portions were most important for perception. It was known, that for the intelligibility (understanding) of speech to the maintained that first two formant peaks ($F1$,$F2$) are most important. There are a number of acoustic parameters in speech, such as fricatives, plosives, consonants, and vowels, all of which as a part of speech inherently contain formant information. It was also felt that the perception of sound is influenced by temporal offsets, (arrival time of signal). In [3], it was postulated that in complex sound environments, delaying the stimulus of interest compared to the competing sounds, made it easier to detect.

This theory is valid in that a mobile phone user could delay their speech relative to the noise, as long as the noise was short in time. Unfortunately, the background noise in most cases is difficult for the user to predict, and if this technique were used, mobile phone conversations would

become very stilted affairs. A more realistic situation is that the speech and noise will coexist in time, making this method less useful. This did not mean that the premise was inapplicable to the mobile phone speech enhancement scenario, as it was the immediate change between the background and the enhanced speech was the basis for the frame nulling technique, implemented in the musical noise technique of Chapter 5. Given that the temporal offset theory was not a valid one for this problem at hand, it was then important to focus on the acoustic parameters of speech, and single out those which are important in speech perception. In [114] page 11, an interesting parallel was drawn between the features of speech production and human perception.

> "Evidence that speech perception is characterised by a process quite different from what is used in the perception of other sounds, is also provided by studies that establish the regions in the brain that play a role in the perception of linguistic and non-linguistic stimuli."

This pointed to the fact that the brain and auditory system had a predisposition for the perception of speech, which was activated on the receipt of speech-like signals. This was also examined in [114] page 13:

> "...listeners and talkers manipulate speech signals in ways that are peculiar to speech. The listener need not be set for speech prior to his hearing the signal; his prepared state is triggered by the presence of a signal that has appropriate acoustic properties. It is postulated that a necessary attribute of this signal, is that it have certain dynamics or time varying properties, among which are the syllabic intensity fluctuations, such as are associated with one of the most fundamental attributes of speech - the vowel-consonant dichotomy."

The implication of this was that the differences in speech between vowels and consonants play a major part in speech perception. By examination of one of these areas, it would be possible to determine how the sex of the speaker affects perception. As there has been some prior work conducted on the recognition of vowels [115], it was felt that this should be the section of speech examined.

6.4.1 Vowel formant position and its effect on perception

Having narrowed the perception system down to looking at vowels, it was important to establish how a vowel was detected in a speech segment. There had to be a defining characteristic of vowels and consonants, which allows humans to identify them individually. In [114] page 21, it was suggested that for vowels the identification depended solely on the formant pattern, and not on the fundamental frequency.

This means that each vowel (and by correlation consonant) has a unique formant pattern, with set relationships between the formant positions. The proposal that fundamental frequency had little importance makes sense, as this allows many different people speaking at separate pitches to produce vowels and consonants, which could be easily perceived by a listener. This did not mean that pitch was unimportant in perception, just that the formant positions defined the speech portion produced, not the pitch of the speech.

Having seen that the formant position was important in vowel production and perception, it then followed that there could be a difference in the positions of the formants $F1$ and $F2$ for the same vowel spoken by people of a different sex. Intuitively it would seem that as females (and children) generally have higher pitched speech than males that this would be reflected in differing $F1$ and $F2$ positions. An example of average formant frequencies for vowels can be found in figure 6.8.

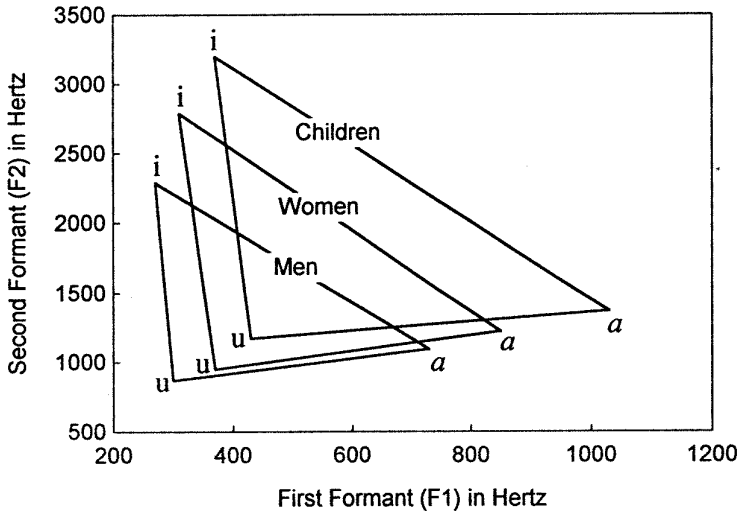


Figure 6.8: *Average formant frequencies for the vowels /i/, /a/, and /u/ as spoken by men women and children as shown in [2].*

It can be seen from figure 6.8 that there is an increase in the $F1$ and $F2$ positions for each of the

three vowels when moving from male to female, and then on to children's speech. Depending on the vowel being spoken, these changes can be quite dramatic, for example the vowel /i/ has an average $F1$, $F2$ of 300 and 2300Hz for a male speaker, and approximately 380 and 2800Hz for female speech. Moving on from female to children's speech, and a similar increase can be seen. Why should this be the case? The formant frequencies $F1$, $F2$ are dependent on the volume behind and in front of the point of tongue elevation respectively (lowering as the volume increases), which means that the smaller the total physical volume, the higher the pitch of speech. The defining element in this case is the length of the vocal tract, which for children will be far shorter than adults, and in general shorter for females than males.

It was previously stated that the first two formants were important in terms of the understanding of speech. What did the differing formant positions mean for the first two formants, when speech was subjected to noise? Was higher pitched speech more resilient to noise, due to the associated shift in the $F1$ and $F2$ frequencies? Jean-Claude Junqua produced some work [116] on listener judgement of speech which had been corrupted in noise. The interesting point about this work was that it examined the possible effects of the sex of the speaker in noise.

The findings of [116] suggested that there was a definite difference in the results obtained from female and male experiments, which resulted from the position of the formants.

> "..depending on the position and amplitude of the second and subsequent formants, which vary in different conditions...the amount of masking induced by the additive noise is more or less important."

The implication here was that dependent on the speech signal of interest, the formant positions could determine how much the speech could be perceived over the background noise. It had been shown how the formant positions for male and female speech differ, which led to the conclusion that male and female speech were perceived differently by the human ear when in identical noise. In order for the results of the CSMT estimation technique to be justified, this would mean that female speech with its higher pitch and subsequently higher formant positions would be easier to perceive, than male speech exposed to the same noise conditions.

In [116] the results obtained seemed to follow this theory where it was stated that:

> "...we observed that when female speakers produced speech in noise their

> second formant was generally over the masking multitalker noise"
> "Using multitalker noise as additive noise we found that there is a strong interaction between the type of noise used and the intelligibility obtained, and that female speakers are more intelligible than male speakers."

It is therefore proposed, that the reason for the differing results for male and female CSMT estimation technique presented are due to the differing ability to perceive male and female speech in noise. The higher $F1$ and $F2$ positions for female make it more resilient to corruption by additive noise, while the converse is true for male speech. Due to this phenomenon, it was harder for the CSMT estimation technique to produce accurate male speech thresholds than female speech. This meant that for the enhancement of the speech spectrum presented, the second formant of the female spectrum was above the main formant of the noise spectrum, meaning that it was possible to reconstruct speech that could be easily understood. For the male spectrum, the lower first and second formants were subjected to a far higher level of interference by the noise, meaning that it was far harder to reduce the effects of the noise on the most important information sections of the spectrum. This led to the dichotomy between the results for male and female speech which were presented here.

6.5 Summary

In this Chapter the technique of estimating the CSMT was examined. It was seen that without a good estimation of the CSMT, there was a good possibility that either important speech information would be removed, or too much residual noise be left in the signal. The problem with attempting to estimate the CSMT, was that there was no access to the clean speech on which to base the estimate. If the noise environment was stationary, this was not too much of a concern. For the non-stationary noise conditions being analysed, this presented a problem, as there had to be a good estimate of the noise to ensure that there was as little corruption of the estimated speech as possible. Any presence of musical noise or spectral nulls would result in thresholds being estimated, which were incorrect.

A new method of CSMT estimation was presented which did not use a SS estimate of the clean speech, but used masking thresholds produced by the corrupted speech and the estimated noise, to try and produce a direct estimate of the CSMT. The relationship between each CB of the signal thresholds was examined to determine the CSMT, given that the corrupted threshold contained elements of both the noise and speech thresholds. By power SS of the thresholds

(not the speech signals themselves), and comparison to the absolute thresholds an estimate of the CSMT was produced. It was seen that for different relationships between the thresholds straight $1:1$ subtraction was sub-optimal and a weighting was required which would alter depending on the dB difference between the CBs of the constituent thresholds. It was seen that four different bands of weighting were required, and any deviation from the ranges explored led to a severe degradation in the ability to estimate the thresholds.

The results of the estimation technique were markedly different for male and female thresholds with female results proving to be by far the most accurate. It could be seen that over the range of tests those which were produced for female speakers were consistently better than the male results. By examining the theory behind formant positions for vowels it was proposed that female speech was easier to perceive in noise than male speech due to the higher fundamental frequency and subsequently higher formant positions present in the spectrum. As these lower formants were the most important in terms of carrying information for speech the less they were subjected to corruption, the easier it became to produce accurate enhanced thresholds and subsequently enhanced speech. It was proposed that this was the reason for the sharp divide in results of the CSMT due to the sex of the speaker.

Chapter 7
The Final Speech Enhancement Results.

7.1 Final algorithm performance

With the implementation of the CSMT estimator achieved, the final perceptual based mobile phone noise reduction/speech enhancement algorithm had been produced. Figure 7.1 shows a flow diagram of the whole algorithm.

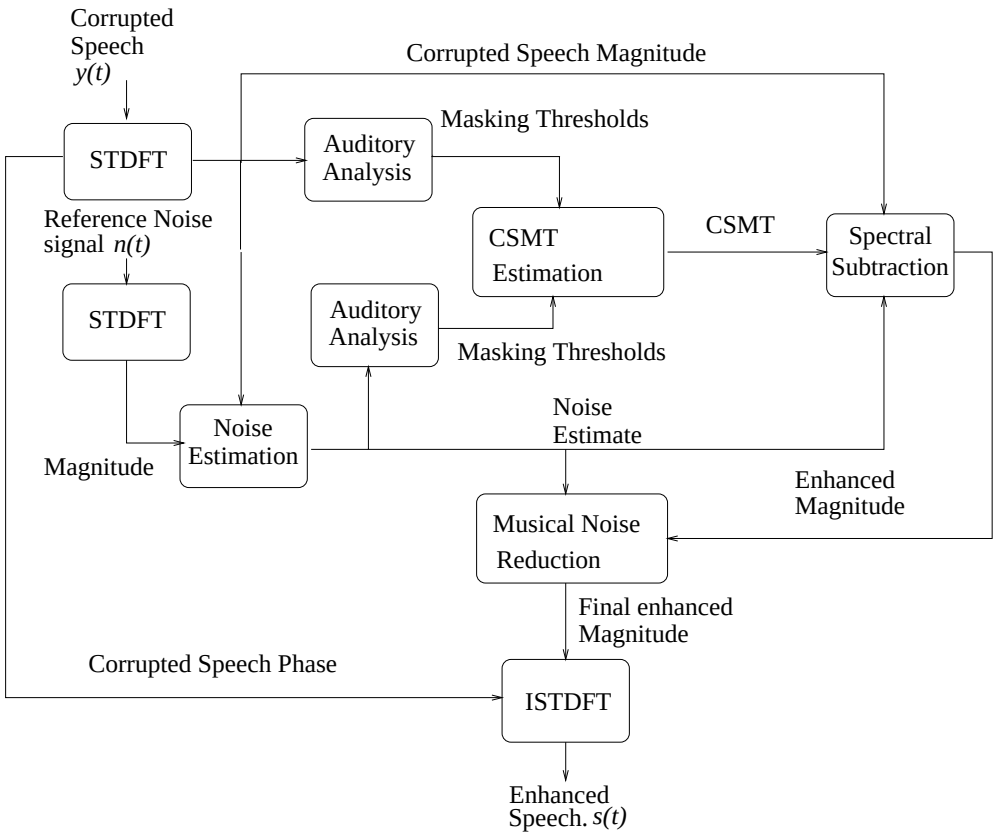


Figure 7.1: *Flow Diagram of the finished speech enhancement algorithm.*

The constituent sections of the algorithm as described in previous chapters are shown.

- SS methodology and the application of auditory analysis (Chapter 3).

- Background noise estimation (Chapter 4).
- Musical noise reduction (Chapter 5).
- CSMT estimation (Chapter 6).

In order to determine the performance of the entire algorithm, a sequence of listening tests were performed. The testing scale used was again taken to be the MOS scale, and the listener was presented with the corrupted signal and the speech after processing. The purpose of this was to determine the quality of the speech which the final algorithm produced, and the MOS testing in this case would show how acceptable the performance was depending on the corrupting speaker and SNR. A group of 20 listeners were used, and each listened to a total of 32 test conditions which spanned four different SNRs and corrupting noises. The average MOS test results for these are given in figures 7.2 and 7.3.

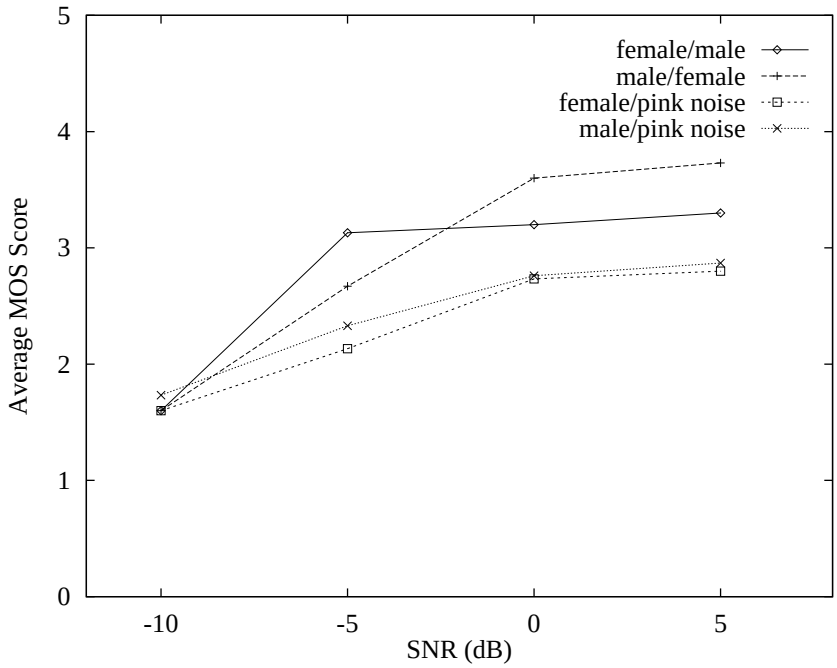


Figure 7.2: *First set of average MOS results*

The results shown in figure 7.2 represent the variation in MOS results for 4 test signals as the SNR increased from -10 dB to 5 dB. The key indicates firstly the speaker which was to be enhanced, and then the signal which was used as corruption. The first scenario examined was corruption by an additive speaker, and by looking at figure 7.2, it can be seen that for both of the cases (female/male and male/female) the average MOS scores increased as the SNR increased.

This was expected, as the less noisy the presented signal, the more likely that the algorithm will be able to approximate the desired speech. The results for these cases lie in between the range *"fair"* to *"good"*, although the most noticeable aspect of these was that the results for the enhancement of the male speech (male/female) outperformed those using the same signal for female enhancement (female/male). Indeed, this was not seen to be an isolated case, as examination of all the MOS results of figures 7.2 and 7.3 showed a difference between the male and female results, with the male case generally producing better results. This was surprising as it seemed to contradict the findings of Chapter 6, where the work pointed to female speech being easier to perceive in noise than male. This was also contradictory to the presented results in section 6.3, where the CSMT estimator was almost 3 times more successful in producing female speech thresholds than male speech thresholds.

For the case of competing speakers, the results for the male were only better than the female case as the SNR moved higher than -5 dB. At this point, a definite plateau became evident in the female results, suggesting that the algorithm as a whole was unable to obtain much more performance for this case. In contrast to this, the results for the pink noise corruption and the tone and music corruption of figure 7.3 show that the male speech was constantly perceived as being of better quality.

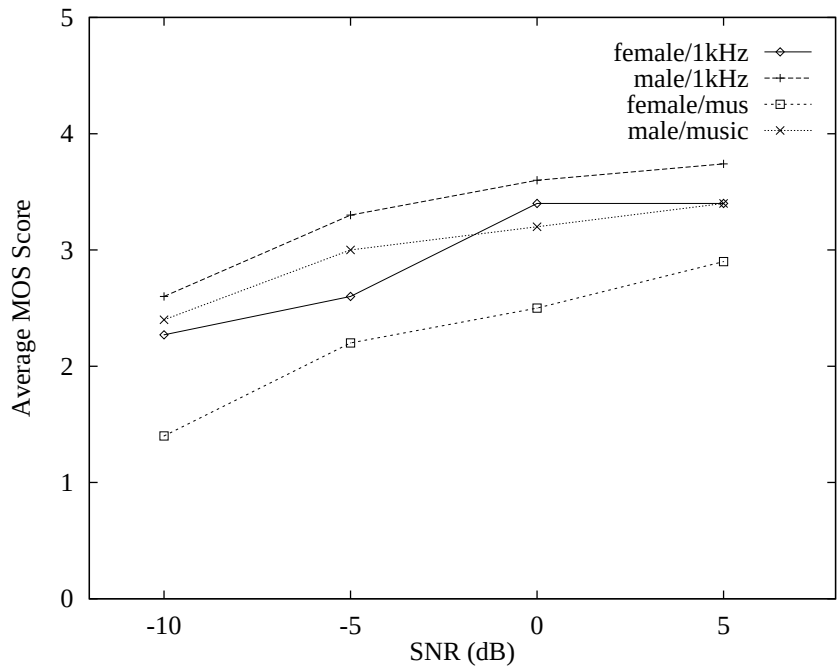


Figure 7.3: *Second set of average MOS results*

It was possible that the results for the male case were prejudiced slightly, by the order in which

the signals were presented in the listening tests. For each test case, the female and then male signals were presented to the listener, and some informal discussions showed that due to this, people could have been influenced by the results they had heard from the previous cases. It would have been interesting to see if a degree of randomisation of the results would have altered these perceived differences. Heuristically, it is proposed that the frequency distribution of the speech segments which were used may be part of the reason for this difference in results. This was shown in figure 7.3, where the results for music corruption showed some of the widest discrepancies between the male and female conditions. The frequency response of the music section used (Brandenburg concerto, 3rd movement) lay more in the region where the important first two formants of the female speech lay. The male voice, on the other hand had much lower first two formants, meaning that they were less corrupted by the presence of the music, and this explains the large variation between the male and female MOS scores. This also had a bearing on the results for the 1kHz corruption, where although the majority of the noise energy was above the formant positions figure F.1(a) shows that there is a significant energy portion below this. The lower formant positions of the male speech lay in a frequency band where the signal of figure F.1(a), had less influence meaning that the information content of the may speech was more robust to the noise.

The overall marking of the MOS testing showed that until the presented test signal reached 0dB SNR, the results were generally felt to be between *"poor"* and *"fair"* which at a first examination was thought to be disappointing. A number of listeners commented that the MOS score would have been higher, except for a *"choppy"* quality which the speech exhibited. Upon further discussion, it was determined that this was due to the spectral nulling portion of the musical noise reduction technique developed in section 5.3.6. Listeners claimed that in some cases, while this allowed for a greater reduction of the musical noise, they found the the loss of some trailing and leading edges of speech portions to be a drawback. If these had been left, listeners claimed that the quality of the signal as a whole would have been improved, beyond any residual noise which may have occurred. This was an interesting observation, as the purpose of the spectral nulling was to make the speech segments easier to hear with regard to the background, although the difficulty in obtaining a flexible threshold value seems to have led to this slight discontinuity which the listeners noticed. It seems that it was not able to achieve the degree of discrimination for the auditory system that was required, and it is debatable whether at this point the spectral nulling achieved was of any use, even although the auditory theory indicated that it should.

As a final test of the potential of the developed speech enhancement algorithm, two tests were run which would establish the robustness of the algorithm. In the previous testing, it was always assumed that the power levels of the noise presented to the reference and primary microphones were the same, it is also only taken that one source of noise would be used as a simulation of a background environment. While pink noise was taken to be a close approximation of the conditions which a mobile phone user may experience, it was decided to take a number of the signals used as background noise, and mix these in unequal amounts, as a simulation of a real life noise condition. This test would be far harder for the finished speech enhancement algorithm to operate on, as it was no longer being operated under the original set of assumptions which were made during the development of the system. This would show how much merit there was in developing this system further.

The test conditions used different proportions of the background elements in the primary and reference signals, meaning that the noise statistics were not similar in the primary and reference signals exposed to the algorithm. In table 7.1, the difference between the percentages of the noise signals can be seen for two test cases.

Background Noise	**Percentage**			
	Primary 1	Reference 1	Primary 2	Reference 2
1kHz Tone	35	10	0	40
Music	55	35	60	0
Pink Noise	10	55	40	60

The difference in segmental SNR for the two test cases was set at 0dB, and the spectrograms presented in figure 7.4 show a portion of speech signal for each noise condition. The noise conditions for the two signals can be seen in figure F.1

Each of the signals represented in figure 7.4 were passed through the final speech enhancement algorithm, and reconstructed to see how robust the technique would be to the different percentages of the background signals. The results of this can be seen in figures 7.5 and 7.6. From the results in figures 7.5 and 7.6 the algorithm managed to reproduce the majority of the speech signal. It can be clearly seen that in the second test case, a greater portion of the speech has been removed than in the first. Overall, given that the percentage levels of the three noise sources had changed considerably between the primary and the reference signals presented to the algorithm, it managed to produce a speech output which upon informal listening,

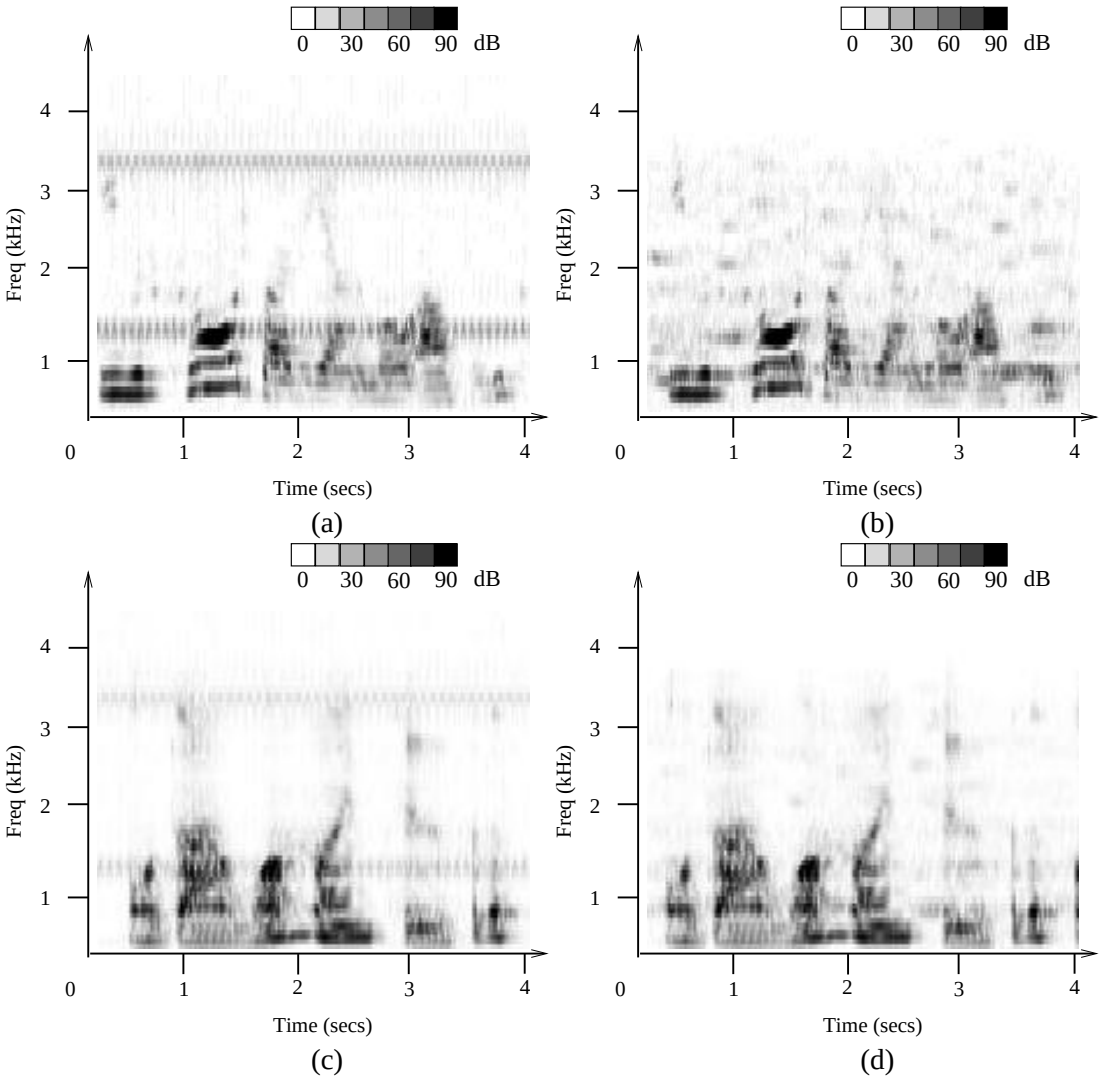


Figure 7.4: *(a)Female speech test signal 1 (b) Female speech test signal 2 (c) Male speech test signal 1 (d) Male speech test signal 2*

was easy to understand. This result was encouraging, as it showed that the developed speech enhancement algorithm was able to cope with different noise conditions being presented in the primary and reference microphones, which is a very likely scenario in real life.

7.2 Summary

The performance of the final speech enhancement algorithm was examined by the use of MOS tests. The results showed that subjectively the output of the proposed speech enhancement algorithm was perceived to be *"fair"*, with regard to the speech quality. It was noted that the

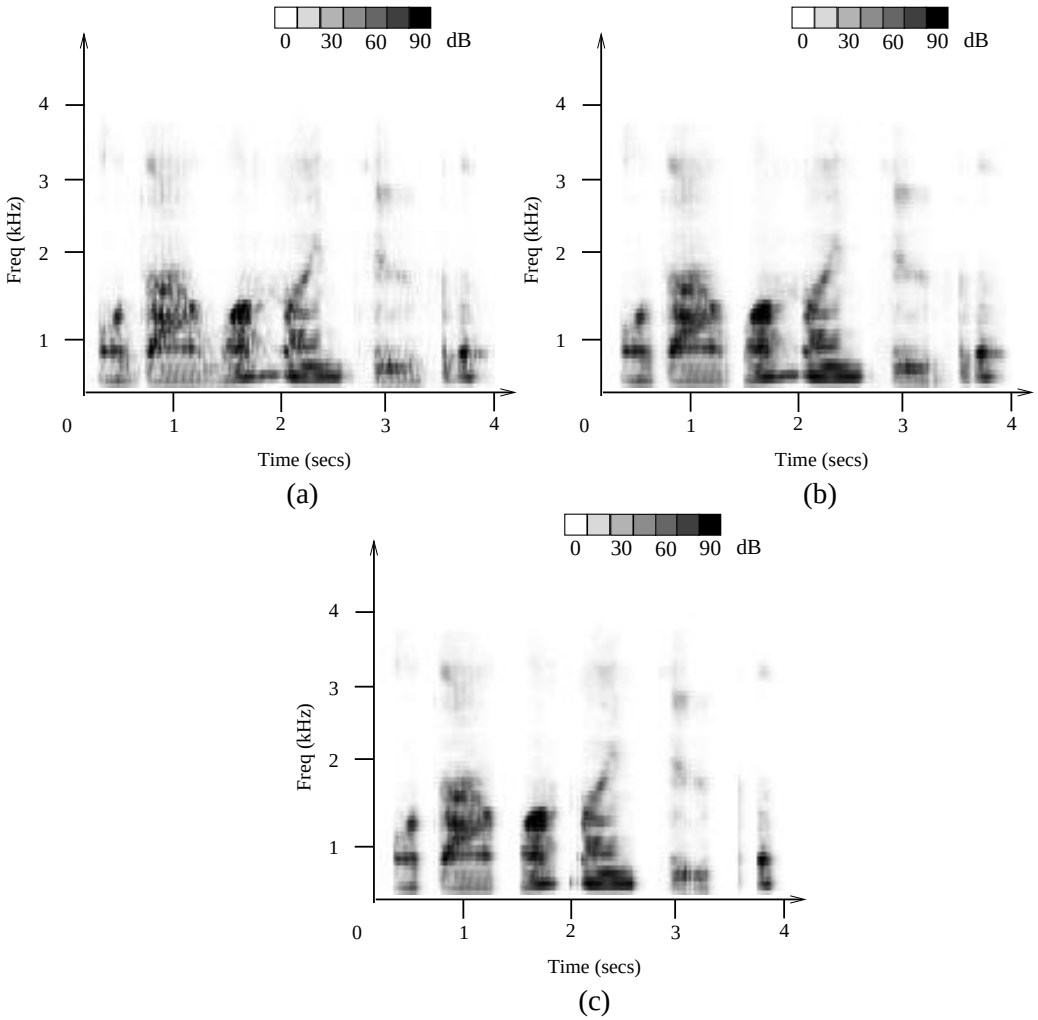


Figure 7.5: *(a)Clean male speech (b) Speech from test signal 1(c) Speech from test signal 2*

quality of the speech in the algorithm was affected by the spectral nulling technique, causing some speech discontinuity which the listeners felt lowered their potential MOS scores for the algorithm. The algorithm was able to remove both stationary and non-stationary noise in these low SNR cases which was encouraging, as the ability to produce results which were understandable at such SNR shows that with improvement of the SNR of the signal, the results produced by the algorithm would improve. A short test was performed, which tested the robustness of the algorithm by removing one of the assumptions of the previous testing. This presented the speech enhancement with primary and reference signals, which varied in the percentile of three noise sources which were present. The resultant speech from this was surprising, in that it was intelligible and easily recognizable, showing that the proposed algorithm does have merit in the area of non-stationary mobile phone speech enhancement.

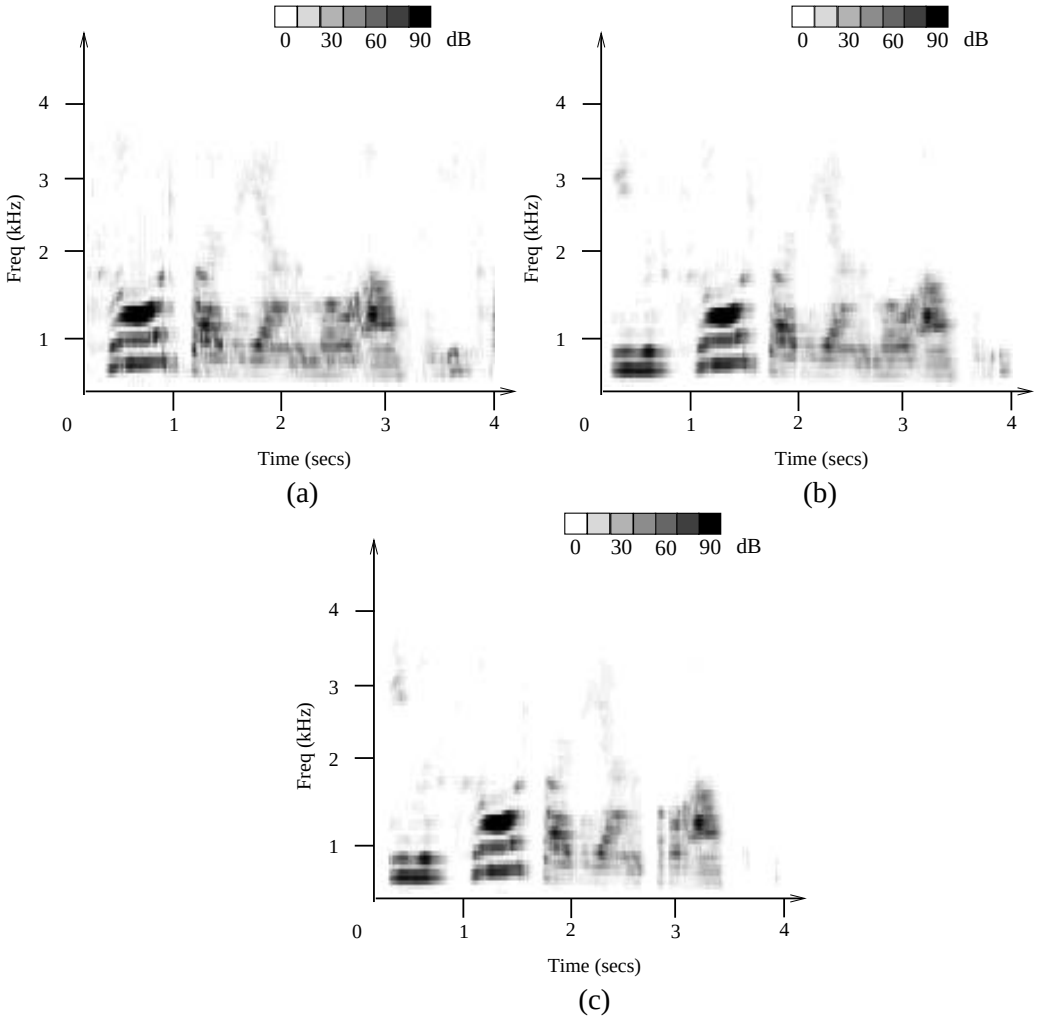


Figure 7.6: *(a)Clean female speech (b) Speech from test signal 1(c) Speech from test signal 2*

Chapter 8
Conclusions.

8.1 Conclusions

Aims

The aim of the work presented here was to produce a perceptually based speech enhancement system for implementation in a mobile phone handset, which could operate in a stationary or non-stationary environment. The algorithm was required to be able to reduce the environmental noise, so that the speaker could be clearly heard and understood. The quality of the speech after enhancement had to be high, as this is a major incentive for subscribers to cellular systems. If the enhancement produces poor quality speech, then consumers would be less likely to choose a handset which used it, as they have a high expectation of the quality of service they should receive. It was just as important that as little corruption of the enhanced speech was produced as possible.

As this algorithm was designed to be a pre-processor system implemented in a mobile handset, it was important to ensure that it was not so unwieldy, that it could not be implemented in such a handset. While a real time solution was unreasonable, it was hoped that in a fully developed system, a number of the software routines could be optimised in hardware, reducing the computational time required. The algorithm should be able to adapt to a variety of conditions, and also cope with a variety of different speakers automatically.

Spectral subtraction and perceptual enhancement

While not all of the criterion were fulfilled, some significant findings were made which showed how an auditory based speech enhancement algorithm could be implemented in a non-stationary environment. The discovery that the use of power SS, which was the most commonly used SS approach for speech enhancement, was not the optimal choice for the non-stationary noise case was a new development. It could be seen that power SS produced more speech corruption than that from the simpler magnitude SS technique. While this seemed a surprising discovery to make, it is worth pointing out that little work had been performed on the effects of non-

stationary noise conditions on SS, and how this would affect enhanced speech which had come from very low SNR signals. The important point about SS, was the relatively high performance which could be achieved for such a low computational complexity, which made it ideal for implementation in this algorithm.

Having seen the benefits SS could have, it was discovered that no account was made in this enhancement technique of the human perceptual system. It was seen that algorithms without perceptual enhancement could produce processed speech, which had artificially boosted regions causing the speech to sound unnatural. By examination of the auditory process, it was seen that the cochlea contributed the most mechanically to noise reduction. By modelling this and introducing it into SS, the enhancement only operated on those portions of the corrupted speech which were perceptually relevant. This could lead to a reduction in the time required for the enhancement, depending on the computational time required to produce the masking thresholds. No work was performed to examine the overhead which the perceptual computation introduced, as it was felt that the improvement in the speech quality which resulted, outweighed the performance increase that may have occurred. The masking thresholds also allowed the noise reduction to produce enhanced speech which was easier for the ear to understand, as an account had been taken of the perceptual interaction of the background noise and the speech.

Noise Estimation

Having chosen the perceptual SS methodology, it was then important to determine how to increase the performance of the SS, without changing the underlying nature of the spectral subtraction. Significant performance boosts could be achieved by presenting the most accurate noise estimate to the SS possible, without resorting to complex noise modelling. An examination of single sensor noise reduction systems showed they suffer from a performance disadvantage, as they could only estimate the background noise during speech pauses. For non-stationary conditions this was a problem, and the choice of a dual sensor approach was proposed, which placed a directional reference microphone on top of the antenna of the handset. This allowed the noise estimation to track the background noise while there was speech present, and allowed adaption if the noise conditions changed. The NEAH technique looked at sections of the signal, and tried to reduce the inaccuracy of the noise estimation by verifying if a frame of the signal was speech or noise. This had the drawback of delaying the estimation by one frame (16ms), but it was felt that the benefits of increasing the estimation accuracy outweighed this delay. MOS listening tests comparing the output of the NEAH technique with other noise reduction

approaches showed, that the noise estimation produced by NEAH resulted in speech of a higher quality with less musical noise artefacts.

In any enhancement algorithm, only one estimation technique is used to determine the noise conditions, and while this was adequate for stationary noise, it was felt that the use of a number of techniques in parallel may result in a more robust noise estimation in non-stationary conditions. The PANE technique used a number of computationally similar noise estimation techniques, along with NEAH. It was hoped that the parallel architecture would reduce erroneously detecting speech as noise and vice-versa. The improvement in SNR achieved, and the results from the subjective MOS tests showed, that the PANE technique performed at least as well as the NEAH, and in many cases outperformed it. The MOS results also showed that listener believed the processed speech to be of a *"fair"* to *"good"* quality, confirming that a parallel architecture was valid for non-stationary background noise.

Musical noise reduction

Even with good noise estimation, the SS still produced a degree of corruption in the processed speech, which was the major drawback of this approach. The trade off places noise reduction against speech corruption, and to ensure high speech quality, musical noise must either be prevented or replaced with a perceptually relevant substitute. Examination of the spectrogram showed it was possible to detect musical noise by energy or variance comparison, and a new technique was developed to deal with musical noise, by a combination of energy and variance detection. By concatenation of these, it was shown that the classification could be enhanced, preventing genuine speech points from being erroneously operated on. Examining the spectrogram points surrounding the POI made it possible to substitute anomalous sections with a perceptually closer alternative, bringing these sections closer to the original.

To replace any removed speech, it was necessary to know the probable spectral value of the original. The spectrogram points around the POI in time and frequency, could be used to help determine whether a section of low spectral values were intended, or were erroneously nulled speech. The best replacement point could be provided by a linear predictor trained on surrounding spectral points. To prevent any compounding of errors a one step forward backward predictor was used, and this technique was shown to replace lost sections of speech. To make the enhanced speech stand out, frames with a low degree of spectral content were artificially nulled. Subjective MOS tests showed that the proposed technique performed better

over a range of signals, than other techniques which have been developed to deal with musical noise. It was discovered there was a discrepancy between the degree of spectral nulling required for female and male speech. This was consistent over both SNR and corrupting signals, and unfortunately, there was not the time to produce a system which could examine the pitch of the speech, allowing the algorithm to maintain the best possible enhancement for the sex of the speaker. This pointed to the possibility that the detection/perception of speech in noise is different according to sex.

Clean speech masking threshold estimation

It was shown, that the application of inaccurate CSMT estimates leads to unpredictable errors in the perceptual SS. If the thresholds were of the wrong shape, then the enhancement could alter the spectral characteristics of the enhanced speech, if they were of the wrong base level, then it was possible to pass more/less corrupted points to the SS, by claiming erroneously that they were audible or inaudible. The typical method of estimating the clean speech, and then calculating the masking thresholds from this was shown to be inherently flawed, as the speech data could be corrupted by energy added from musical noise, resulting in masking thresholds which were nothing like the originals.

To obtain a good CSMT, the signal must have little corruption, and as the two signals present were the corrupted speech and the estimated noise, there was the need for a different way to produce the thresholds. By calculating a set of masking thresholds for the corrupted speech and the background noise, and by examining the relationship between these in each CB, it was possible to determine where the influence of the speech masking thresholds was present. Through power SS of the thresholds themselves (not the associated magnitude spectrum), a reasonably accurate estimate of the CSMT could be produced. After application of this technique to a number of different signals, it was shown that the technique was a valid alternative to the SS based approach. The experiments showed results that were again different for male and female speech, and to obtain a close female estimation, the algorithm parameters had to be altered in such a way that the male performance was diminished. Even after a great deal of experimentation with the weighting factors, it was not possible to obtain an optimal set for both male and female speech. For each test signal, manual inputting of the optimal male and female weighting factors was required. This difference had not been anticipated, and was a disappointment although it alluded to the same discrepancies shown by the results of the musical noise reduction technique.

By examining the theory on speech production and perception, it was possible to establish there is a difference in the formant positions for male and female speech. While the relationship between formants for vowels and consonants stays constant, the differing fundamental frequency of the sexes places the formants in different frequency regions. The formant positions for male, female and children's spoken vowels showed how the first and second formants differed due to the length of vocal tract, and work in this area also alluded to a difference in the perception of male and female speech in noise, due to the position of the formants in relation to the influence of the noise. It was concluded, that female speech was more easily perceived in noise, due to the higher fundamental frequency and subsequently higher formant positions. This explained the results of the CSMT estimation technique, and the problem of the frame nulling in the musical noise reduction technique, where the higher pitch and breathiness of female speech made it harder to determine frames where there was a low speech content.

Final testing

Tests performed on the completed speech enhancement algorithm showed that it did not perform as well as expected, given the results of each individual section. The MOS testing showed that the completed algorithm was fairly successful in the enhancement of the speech, without too much introduction of speech distortion, apart from at very low SNRs (-10dB). While such a condition was unlikely to be experienced, it helped to show the improvement of the system as the SNR rose to more realistic levels. The listeners did express a preference that the *"choppy"* nature of some test sections be removed, and this would lead to increased MOS scores, as it decreased the overall quality of the output. This was due to the inability to implement an adaptive thresholding for the spectral nulling in the musical noise reduction technique. While it could completely remove all of the residual noise in pauses, the drawback was a loss of some speech leading and trailing edges. Listeners claimed that with these sections present, the improvement in quality would outweigh any noise present during the pauses. An effective speech/pause detector would allow the nulling algorithm to be implemented with minimal impact on the speech information. The MOS testing also showed a perceived difference in performance between male and female speech, something which had been hinted at in Chapter 6. Unlike the work of Chapter 6 though, the male speech was deemed to give the better quality enhancement results. It was felt that some of this may be due to the arrangement of the listening tests, (where randomisation of the presented sequences may have eliminated the male bias), and it could only be theorised that these results could be due to the frequency distribution of the signals used

for experimentation. Results from pink noise experimentation, (figure 7.2), showed a situation where the performance for both sexes was more equal. The robustness test surprisingly showed the algorithm was able to function in conditions where the noise presented to the primary and reference microphones, were significantly different. This revealed that a perceptually based SS algorithm has the ability to perform in a non-stationary environment, but further work is required, in order to ensure equal performance across the range of possible scenarios.

8.2 Future Work

There are a number of sections which could be improved, in order to derive a system which can provide better performance, in terms of noise reduction and speech quality, they are;

- The introduction of pitch detection, in order to determine the sex of the speaker, allowing the algorithm to cater for the differing conditions presented by male, and female speech.
- A study into the effect of the position of the secondary microphone, and experimentation with data collected from an anechoic chamber, using the proposed configuration.
- Further study on the perception of speech in noise to optimise the CSMT and frame nulling sections of the musical noise reduction technique.
- The ability to store multiple noise configurations in memory, and switch between the parameters when the algorithm recognises them.
- The use of a linear predictor to remove musical noise, as well as replace nulled speech sections.
- Applying signals with significantly differing powers, as well as statistical distributions, to the algorithm.
- Examination of the computational power and runtime required for the developed system, and how this could be reduced.
- The rewriting of the software developed, to produce a more compact realisation, allowing the functionality of the algorithm to be implemented in hardware.

Appendix A
Spectral Subtraction Spectrograms

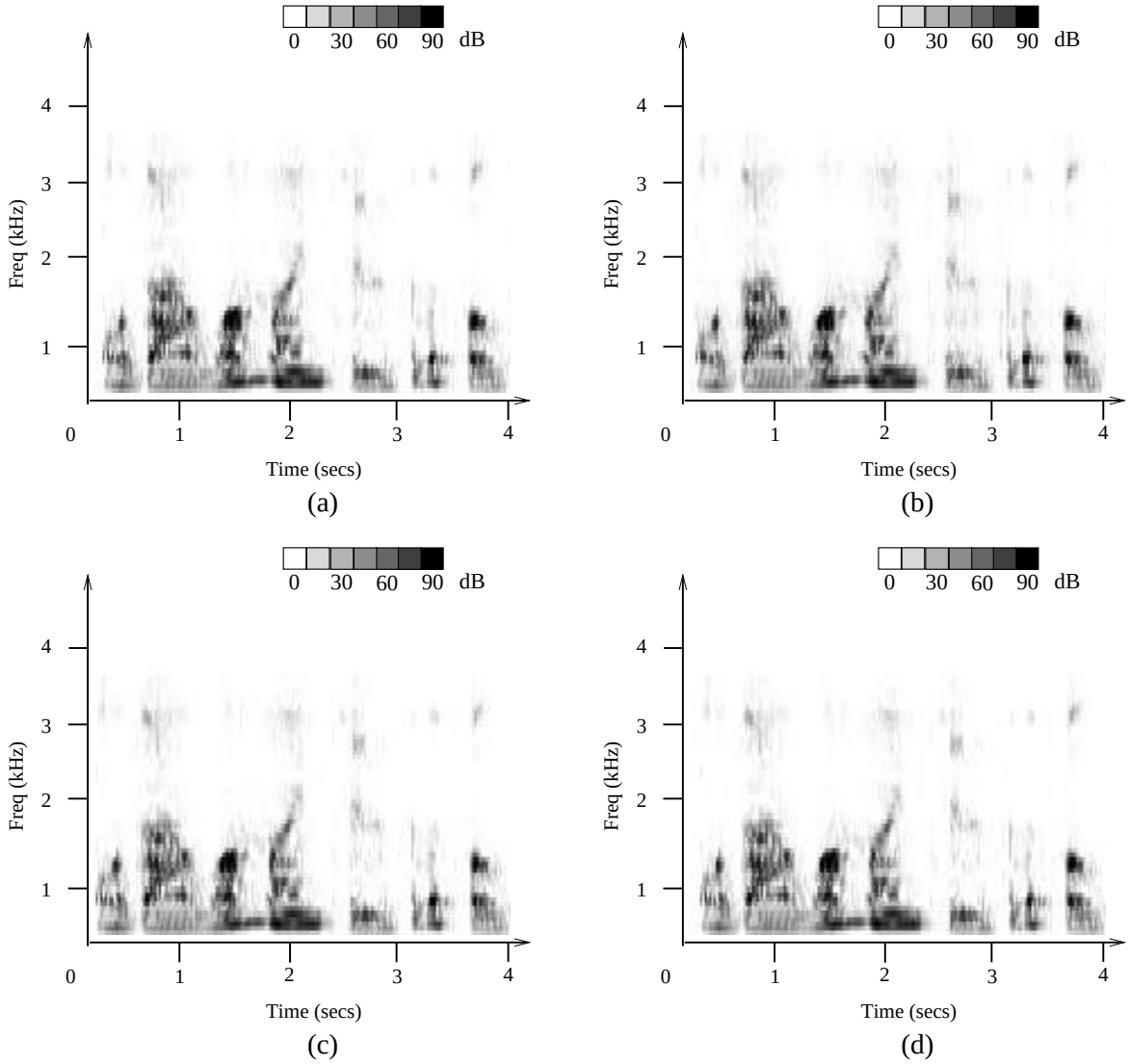


Figure A.1: *Male Voice with (a) Half Wave rectification (b) Full Wave rectification (c)* $\eta = 0.001$ *(d)* $\eta = 0.003$

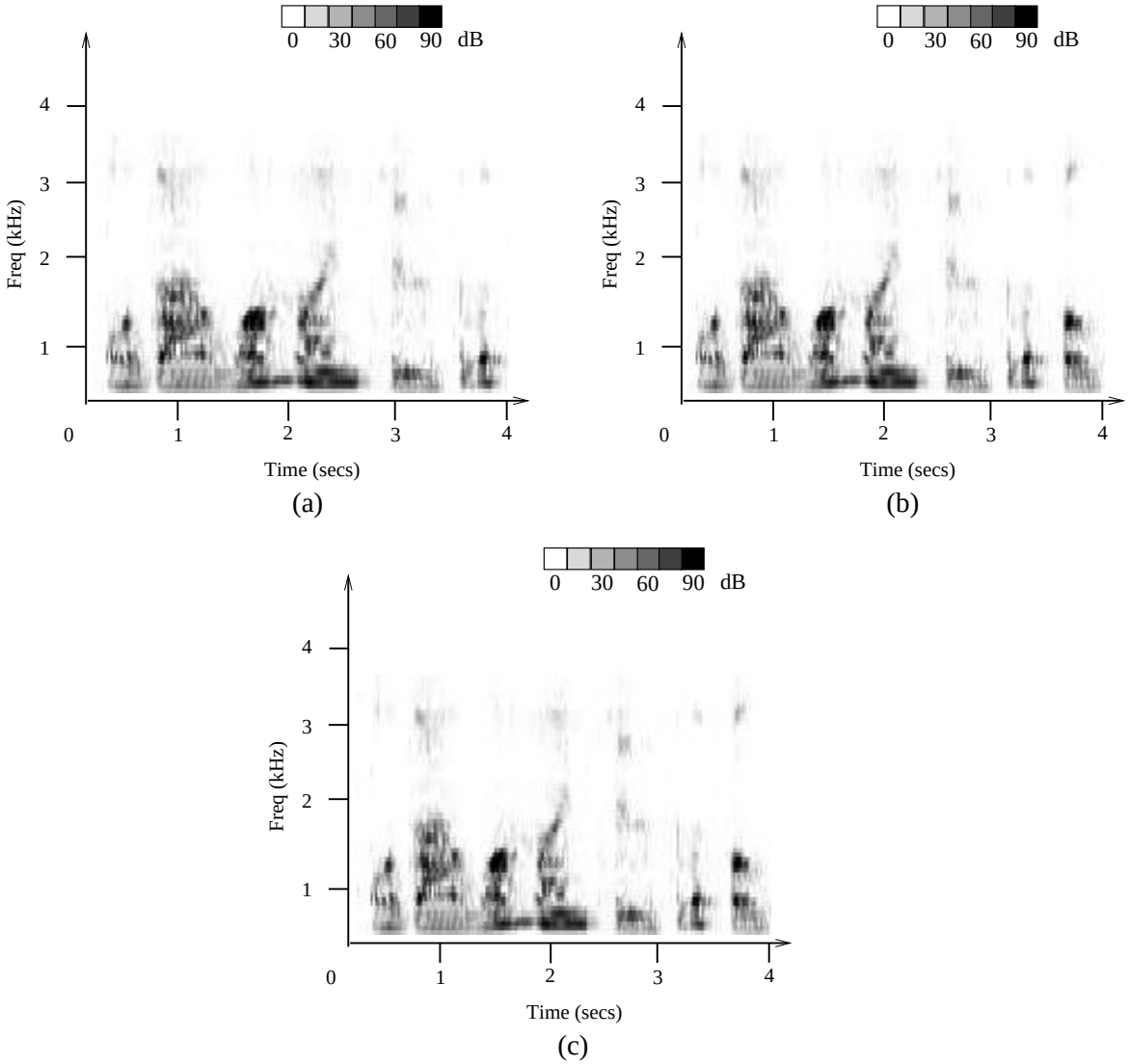


Figure A.2: *Male Voice with (a)* $\eta = 0.008$ *(b)* $\eta = 0.015$ *(c)* $\eta = 0.02$

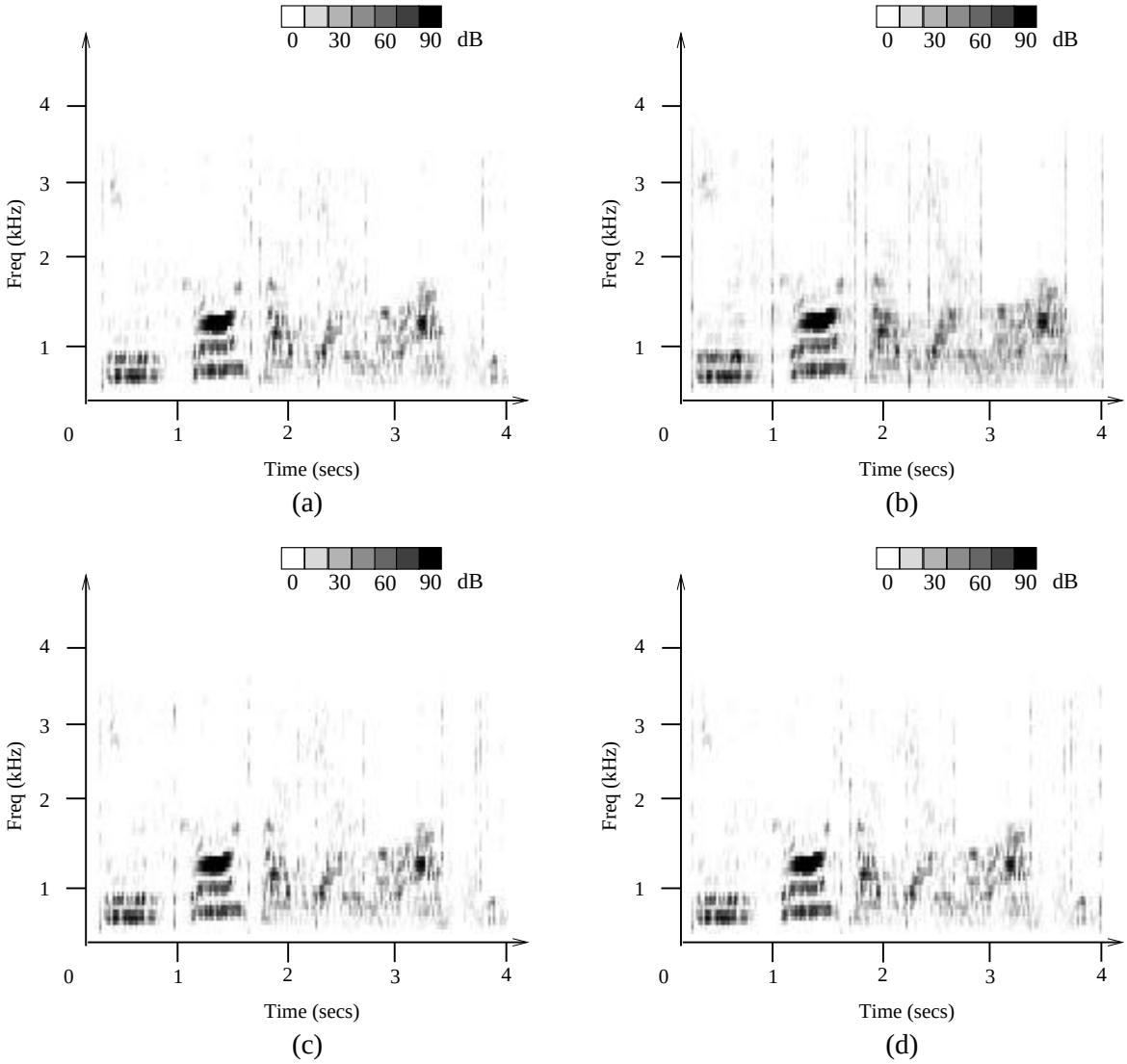


Figure A.3: *Female Voice with (a) Half Wave rectification (b) Full Wave rectification (c)* $\eta = 0.001$ *(d)* $\eta = 0.003$

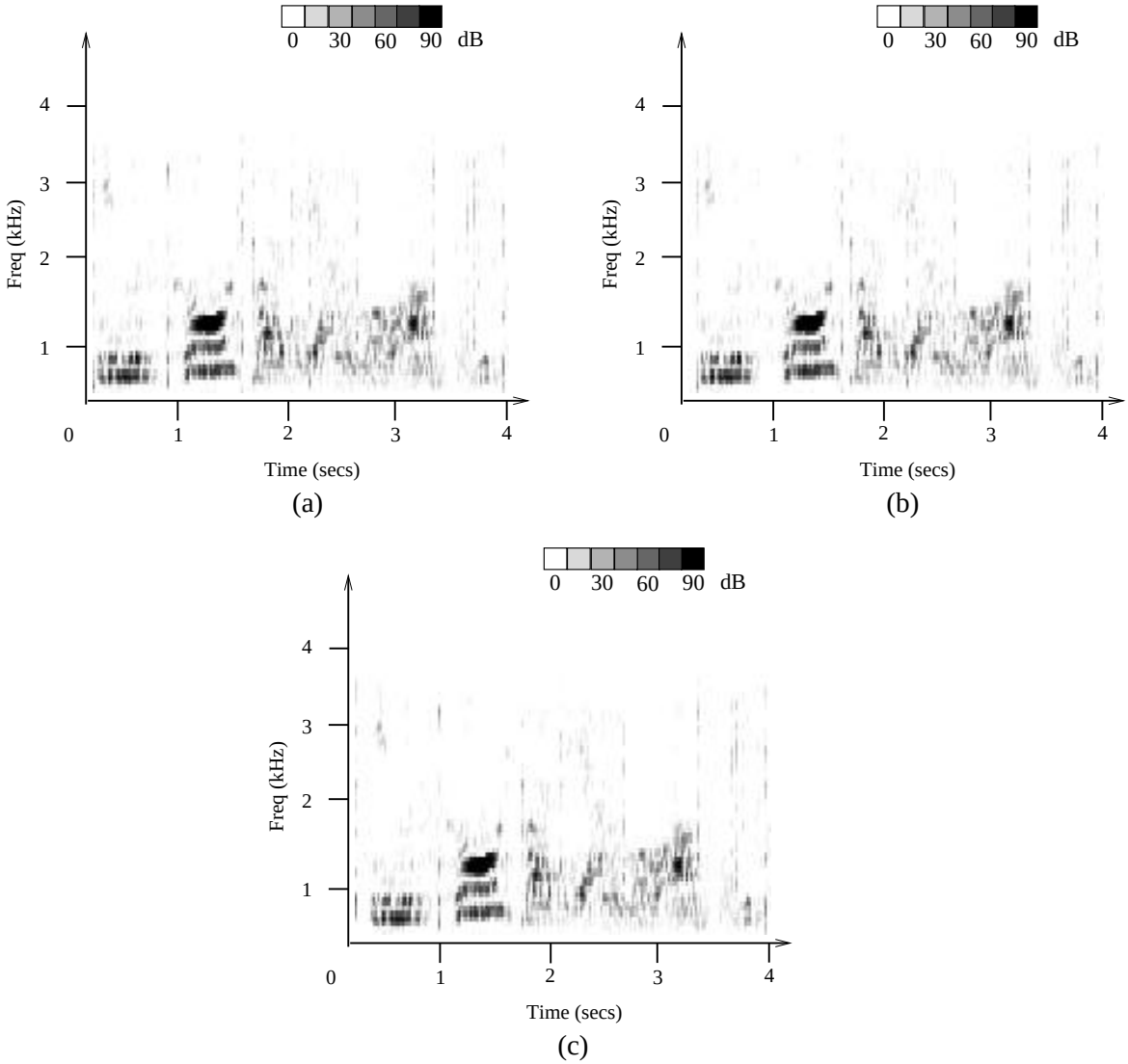


Figure A.4: *Female Voice with (a)* $\eta = 0.008$ *(b)* $\eta = 0.015$ *(c)* $\eta = 0.02$

Appendix B
Spectral Subtraction Spectrograms

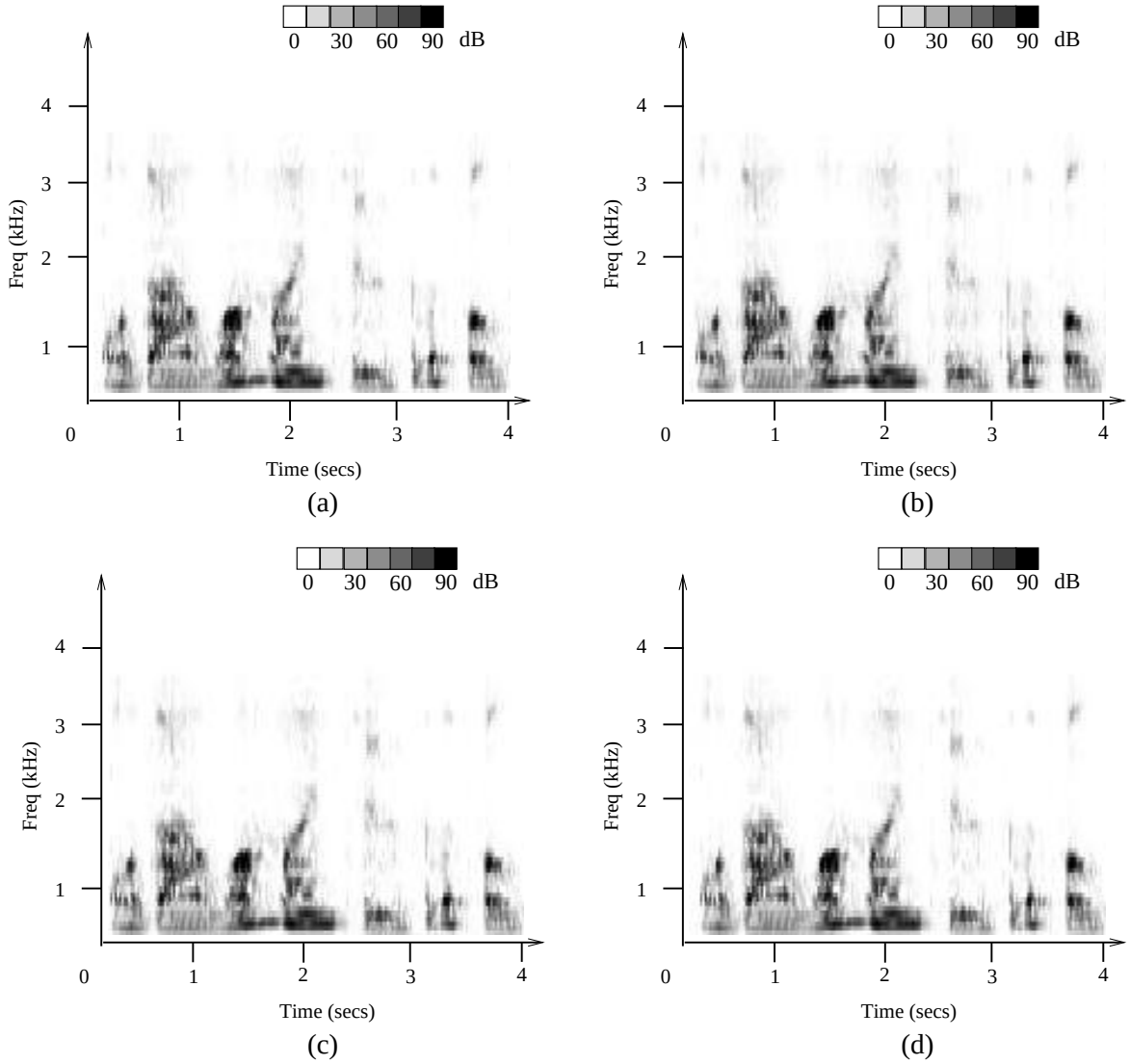


Figure B.1: *Male Voice with (a) Half Wave rectification (b) Full Wave rectification (c)* $\eta = 0.001$ *(d)* $\eta = 0.003$

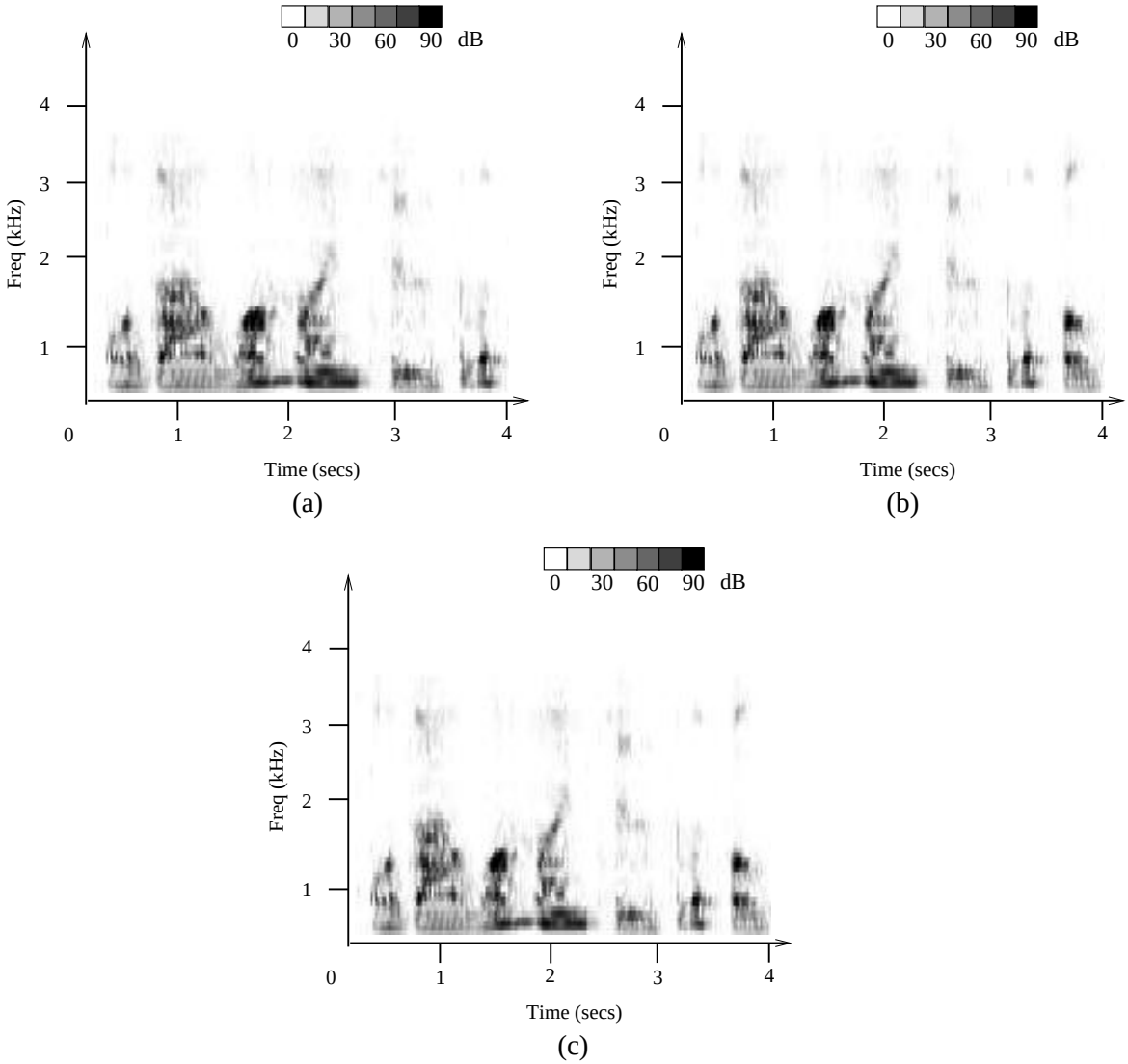


Figure B.2: *Male Voice with (a)* $\eta = 0.008$ *(b)* $\eta = 0.015$ *(c)* $\eta = 0.02$

Appendix C
PANE Algorithm Spectrograms

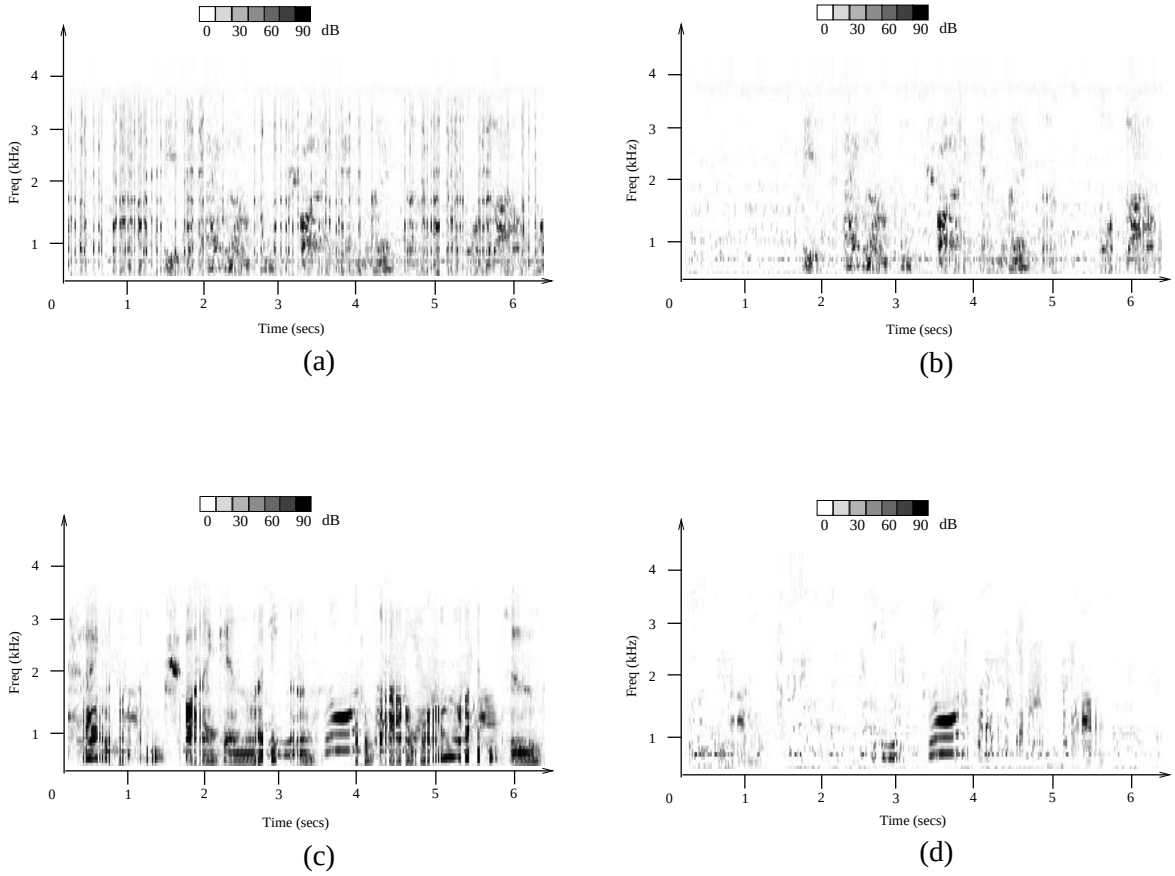


Figure C.1: *Spectrograms sample of PANE algorithm with (a) Male/Pink noise* $-10dB$ *with MIN decision (b) MAX decision (c) Female/Male* $-10dB$ *with MIN decision (d) MAX decision*

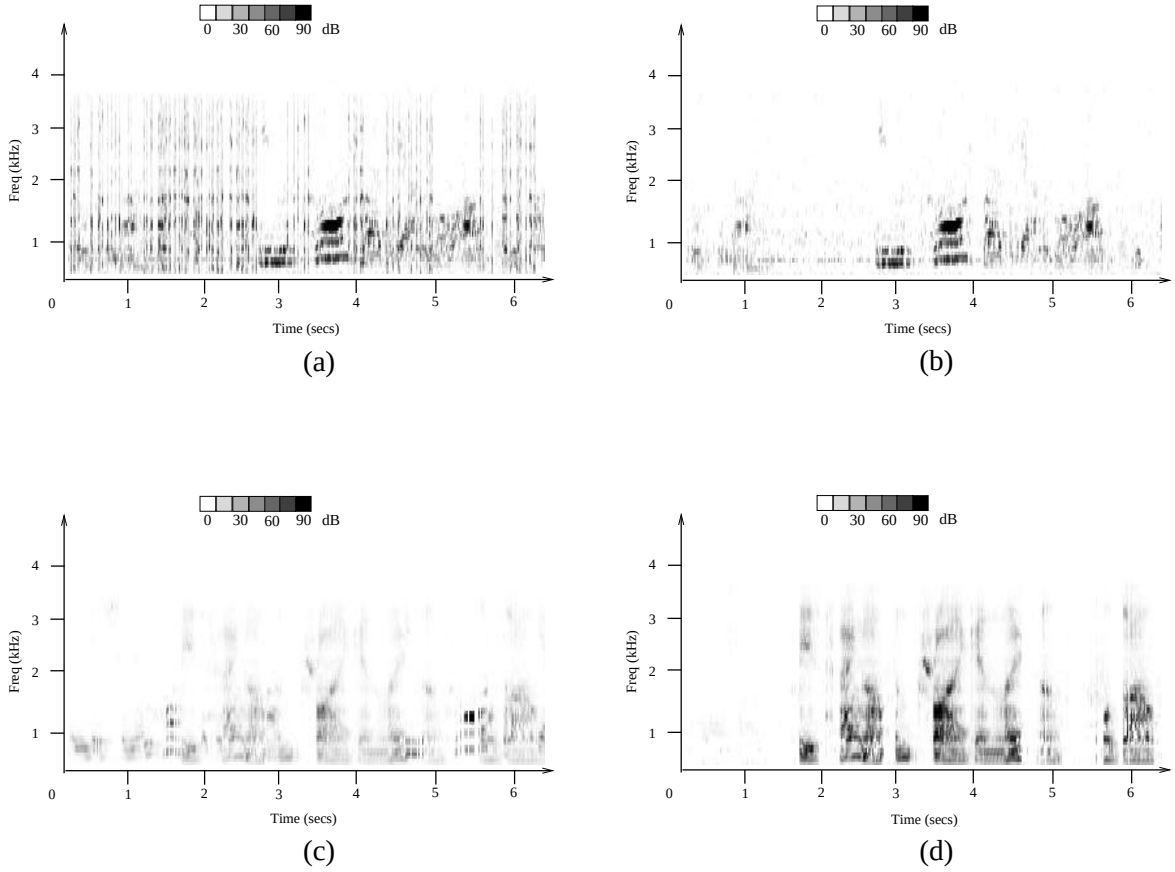


Figure C.2: *Spectrograms sample of PANE algorithm with (a) Female/Pink Noise -5dB with MIN decision (b) MAX decision (c) Male/Female -5dB with MIN decision (d) MAX decision.*

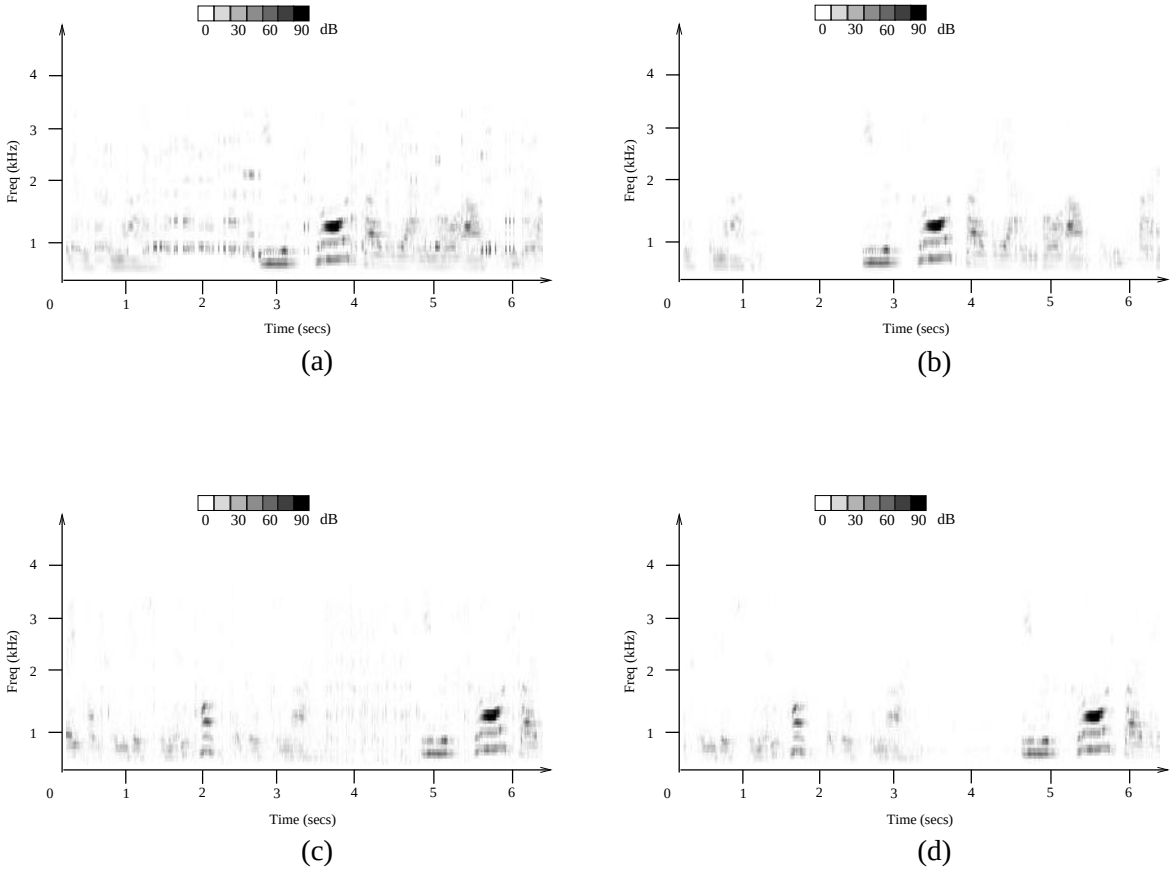


Figure C.3: *Spectrograms sample of PANE algorithm with (a) Female/Music* 0*dB with MIN decision (b) MAX decision (c) Male/*1*kHz tone* 0*dB with MIN decision (d) MAX decision*

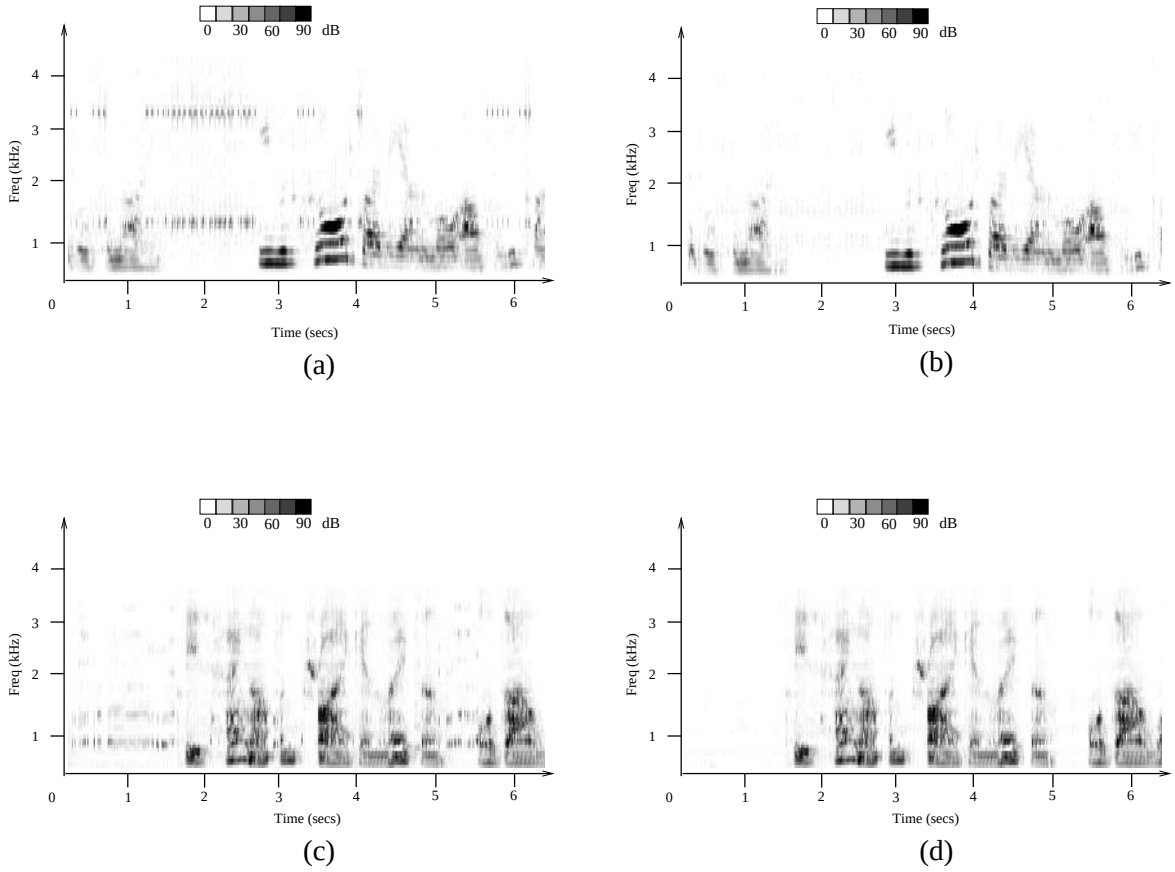


Figure C.4: *Spectrograms sample of PANE algorithm with (a) Female/1kHz tone 5dB with MIN decision (b) MAX decision (c) Male/Music 5dB with MIN decision (d) MAX decision.*

Appendix D
Musical Noise Reduction Frame Null Spectrograms

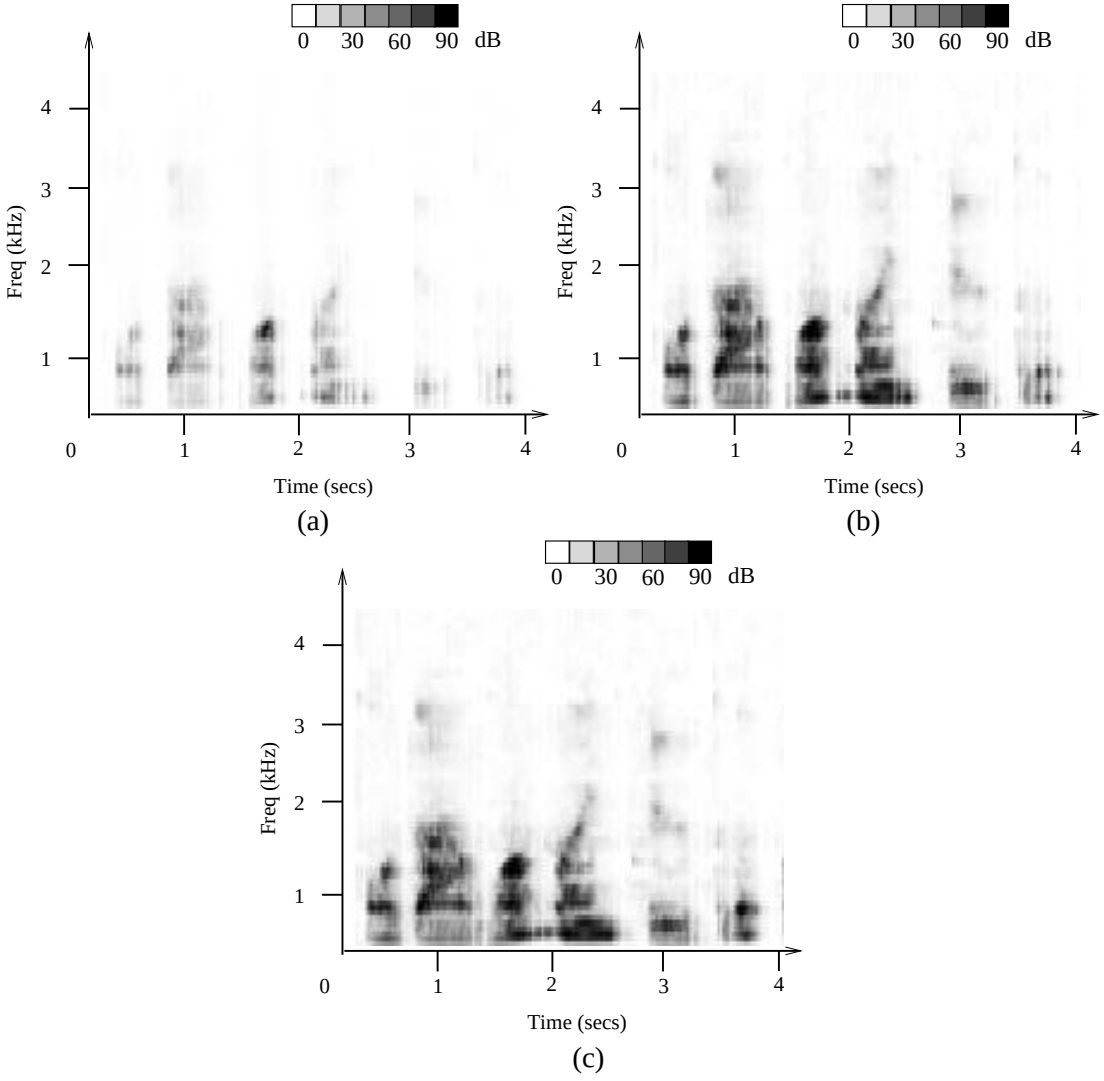


Figure D.1: *Male Speech with Threshold equal to (a) 10 (b) 20 (c) 30*

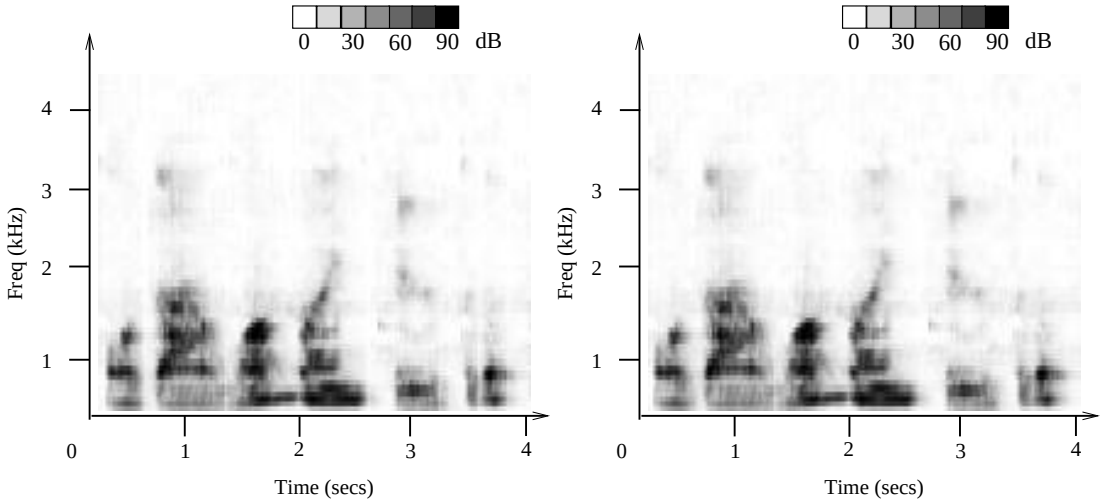


Figure D.2: *Male Speech with Threshold equal to (a) 40 (b) 50*

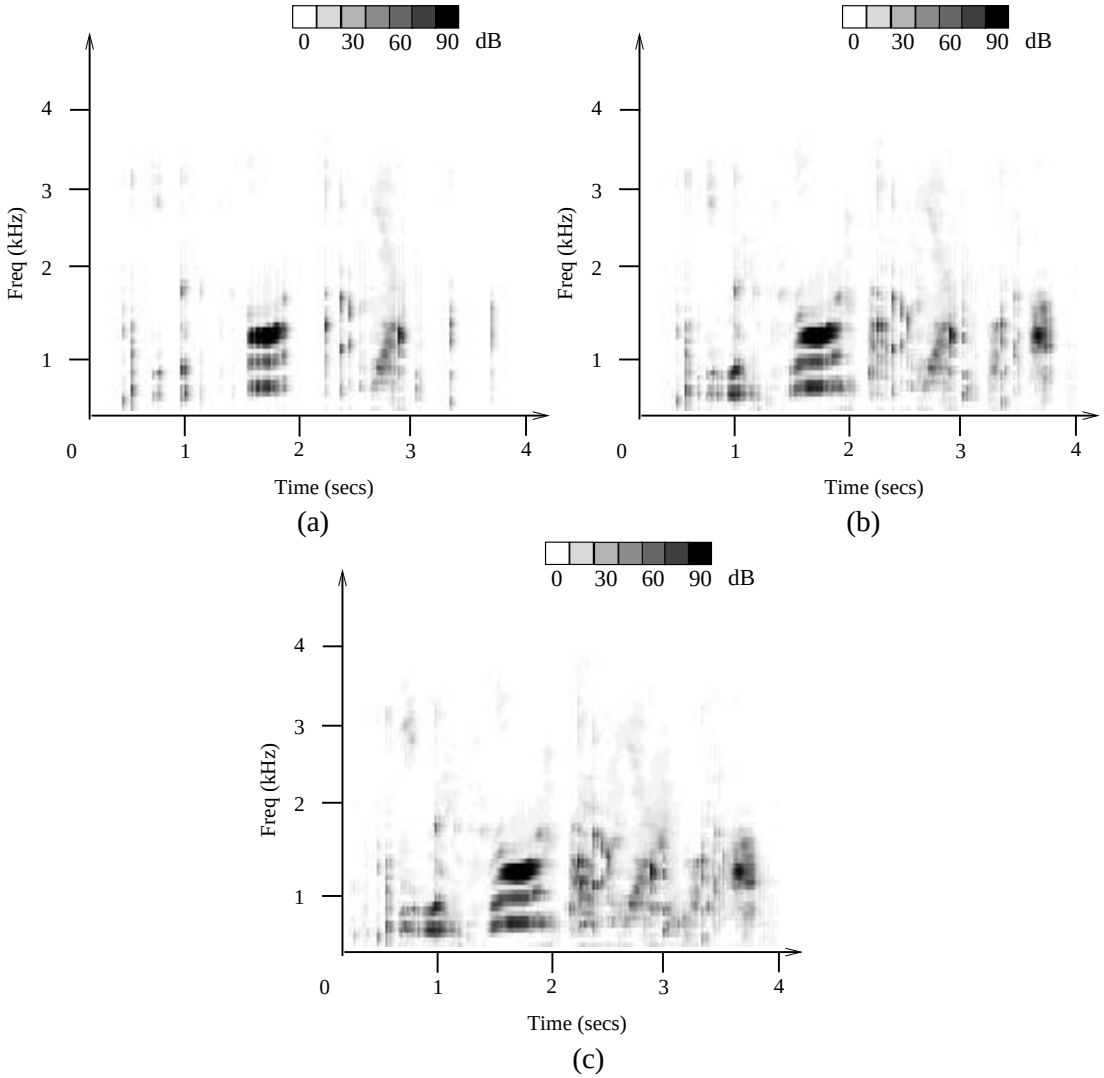


Figure D.3: *Female Speech with Threshold equal to (a) 10 (b) 20 (c) 30*

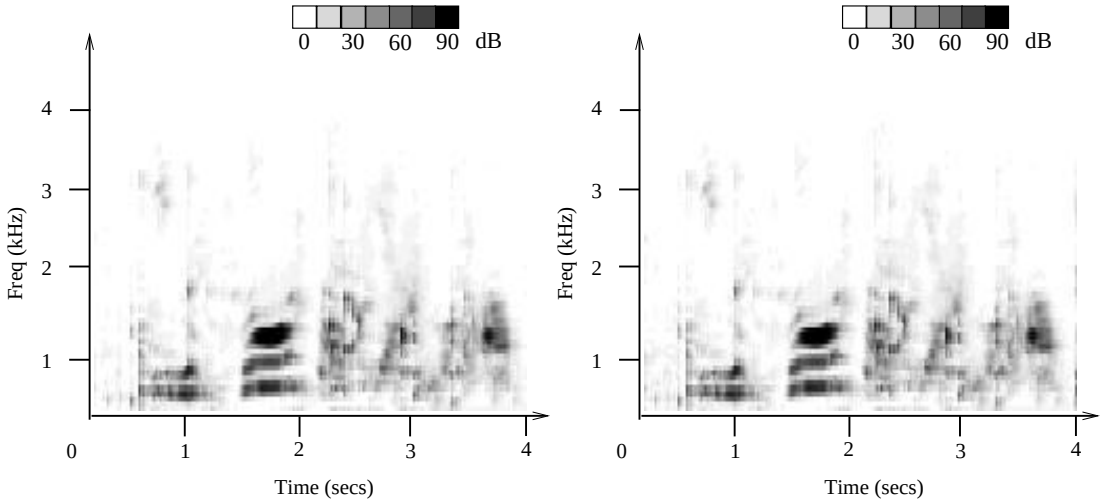


Figure D.4: *Female Speech with Threshold equal to (a) 40 (b) 50*

Appendix E
Clean Speech Masking Threshold estimation results.

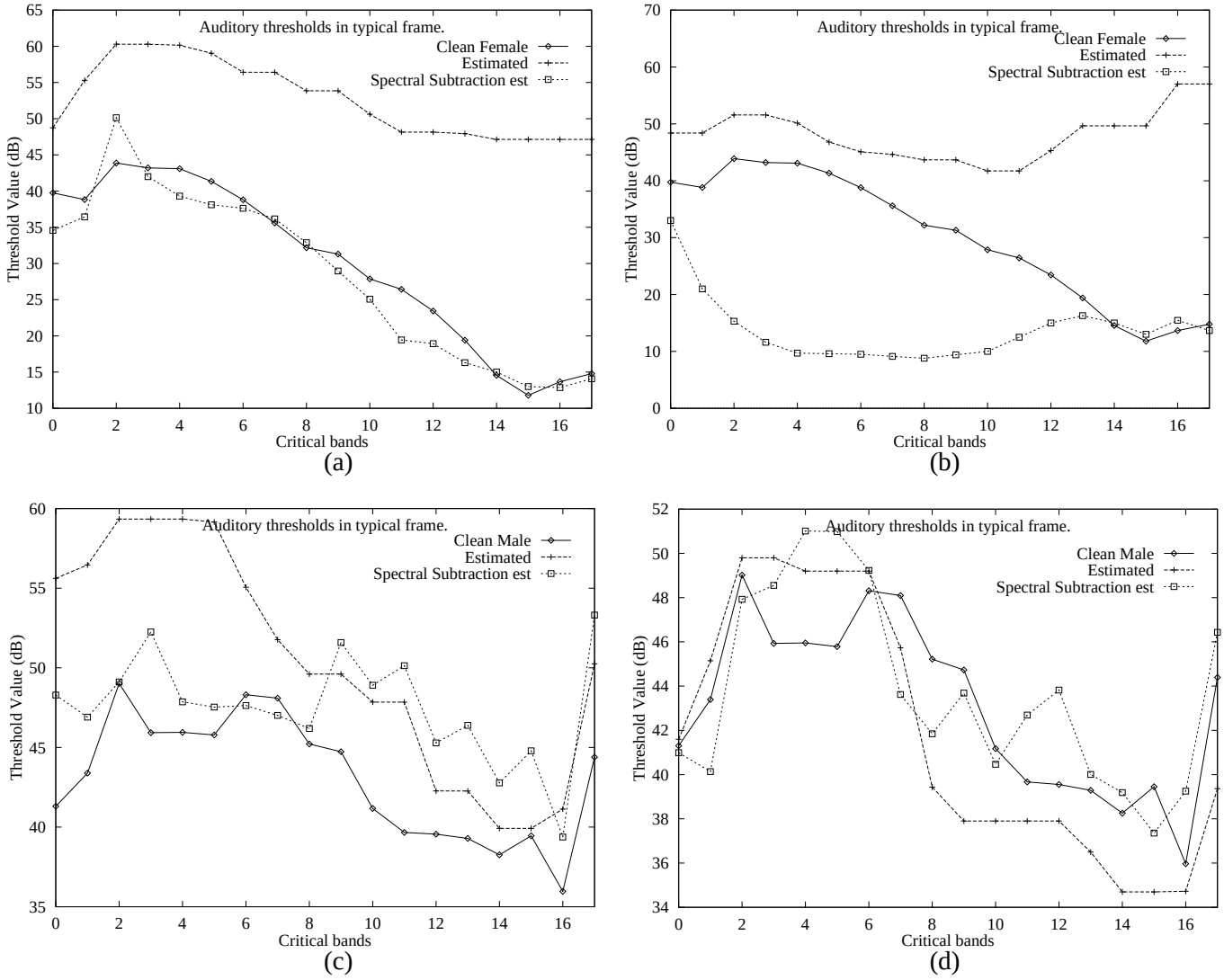


Figure E.1: *Estimate of thresholds for (a) Female corrupted by male speech (b) Female corrupted by* $1kHz$ *tone (c) Male corrupted by pink noise (d) Male corrupted by music all at* $-10dB$

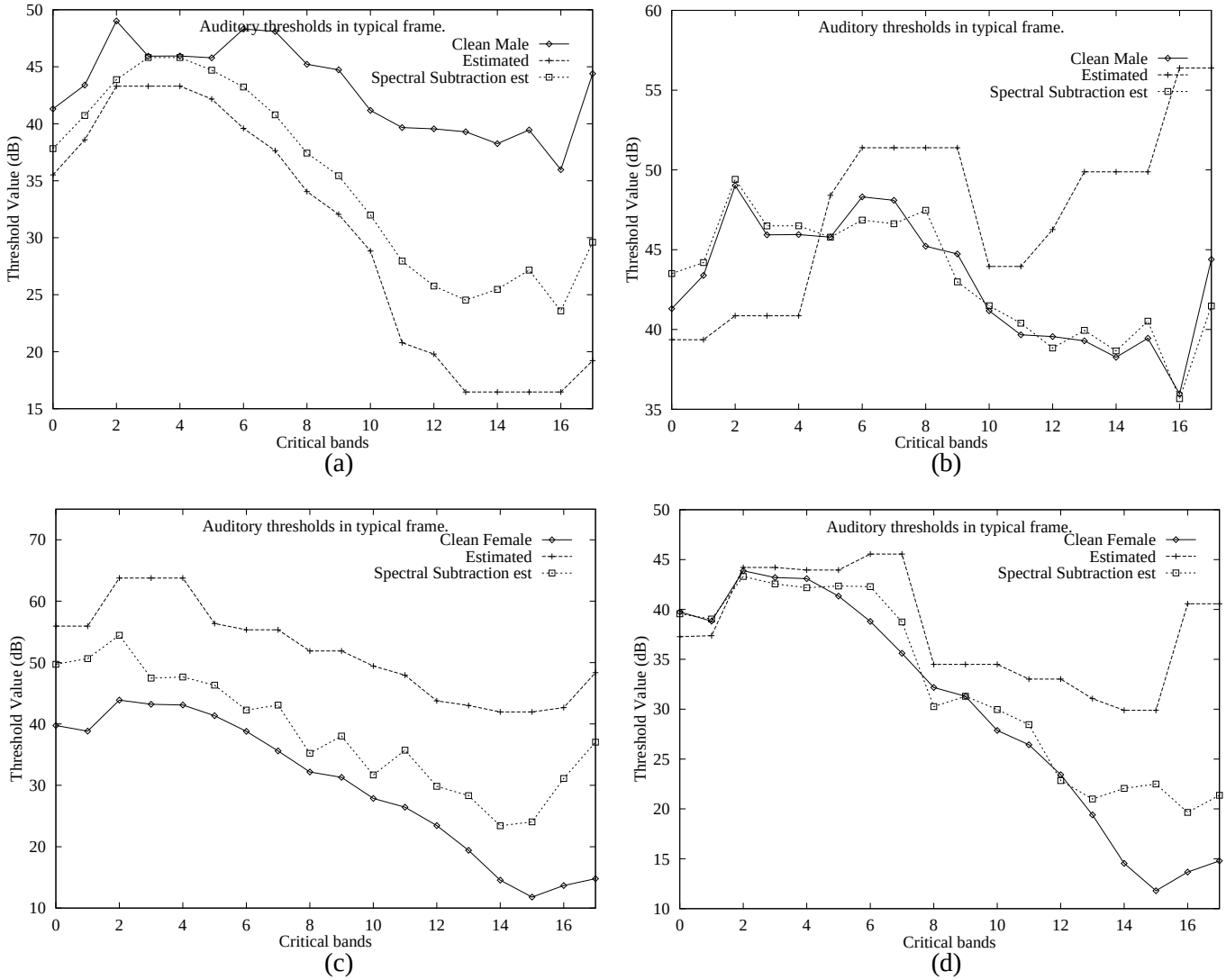


Figure E.2: *Estimate of thresholds for (a) Male corrupted by male speech (b) Male corrupted by* $1kHz$ *tone (c) Female corrupted by pink noise (d) Female corrupted by music all at* $-5dB$

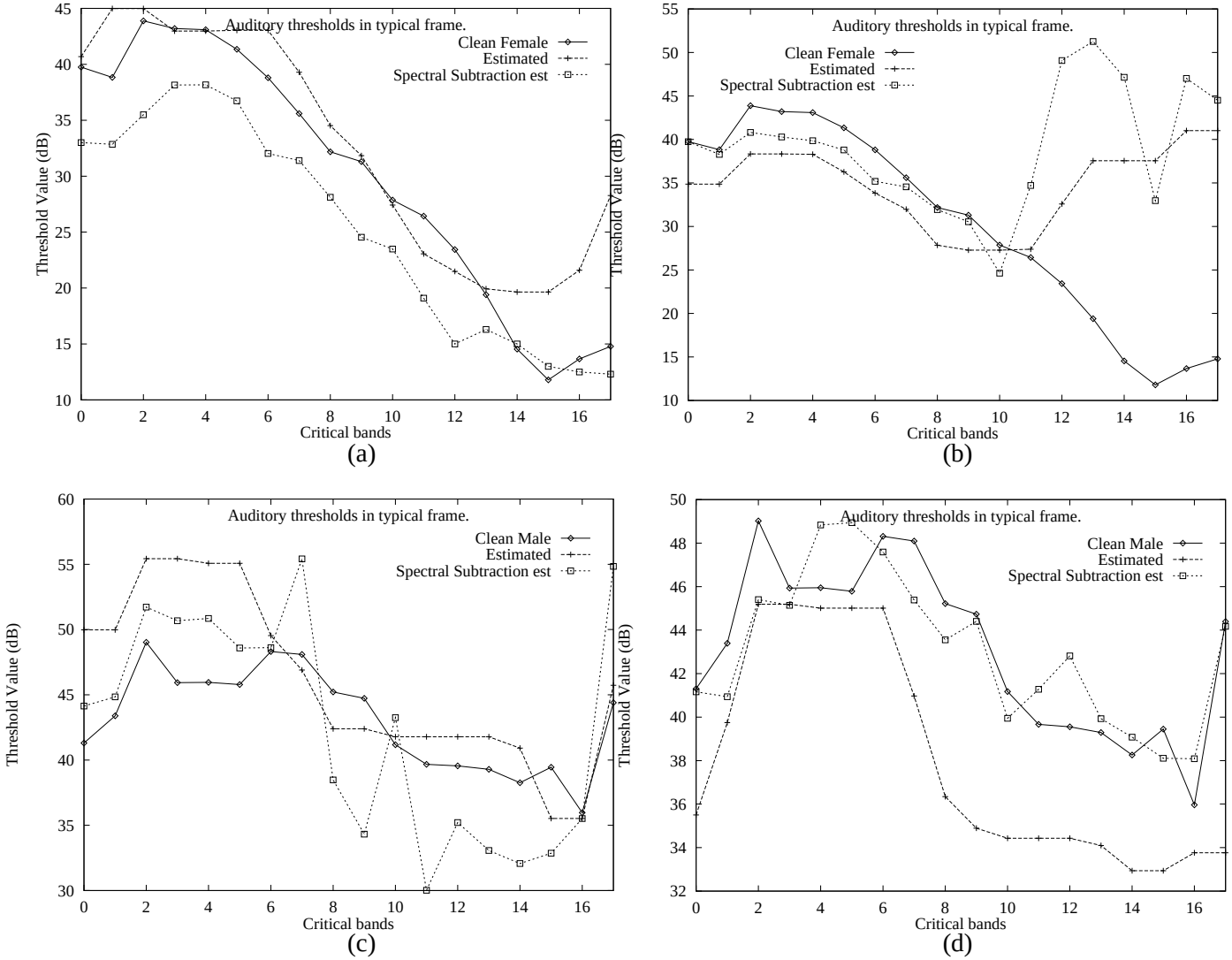


Figure E.3: *Estimate of thresholds for (a) Female corrupted by male speech (b) Female corrupted by* $1kHz$ *tone (c) Male corrupted by pink noise (d) Male corrupted by music all at* $0dB$

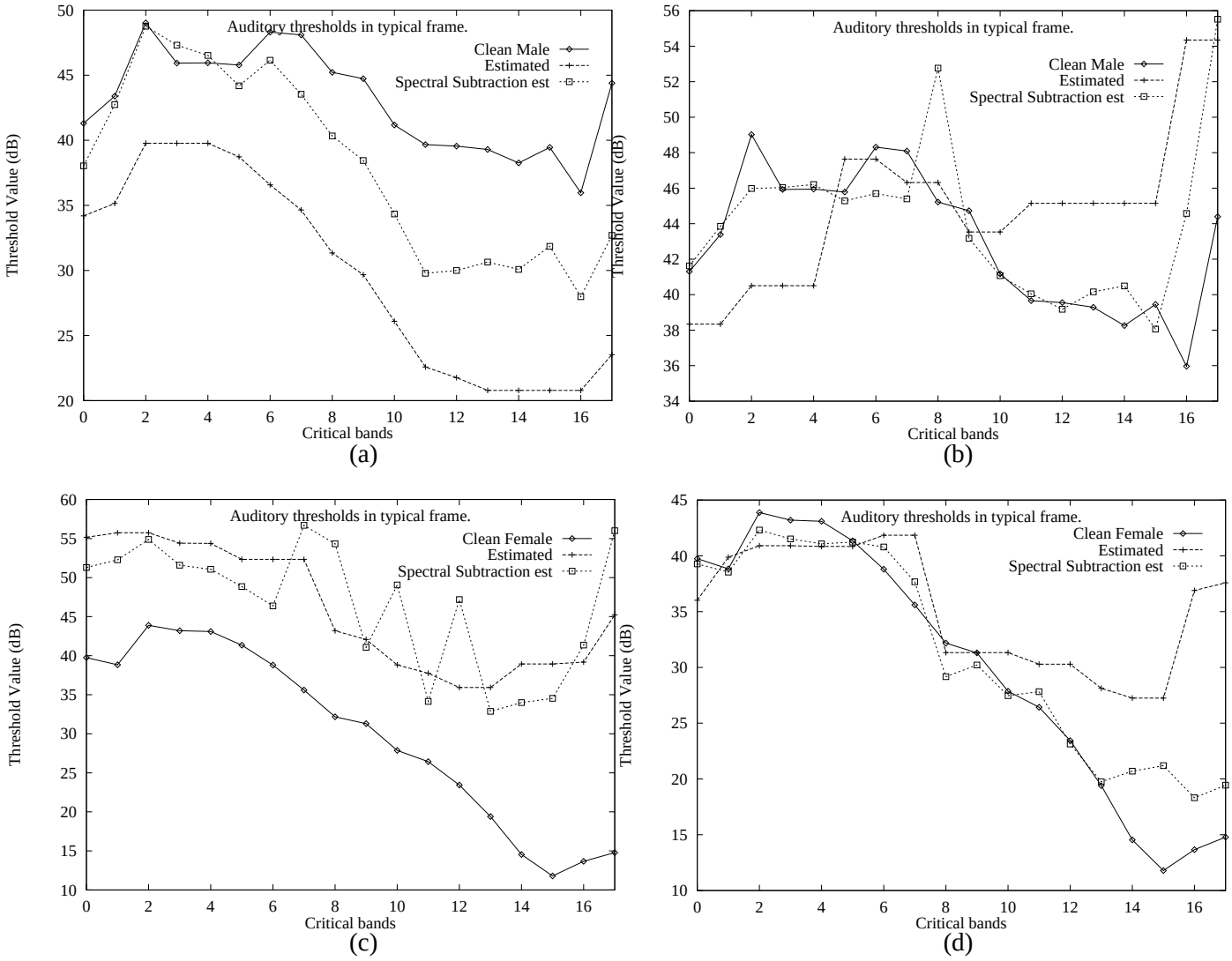


Figure E.4: *Estimate of thresholds for (a) Male corrupted by female speech (b) Male corrupted by* $1kHz$ *tone (c) Female corrupted by pink noise (d) Female corrupted by music all at* $0dB$

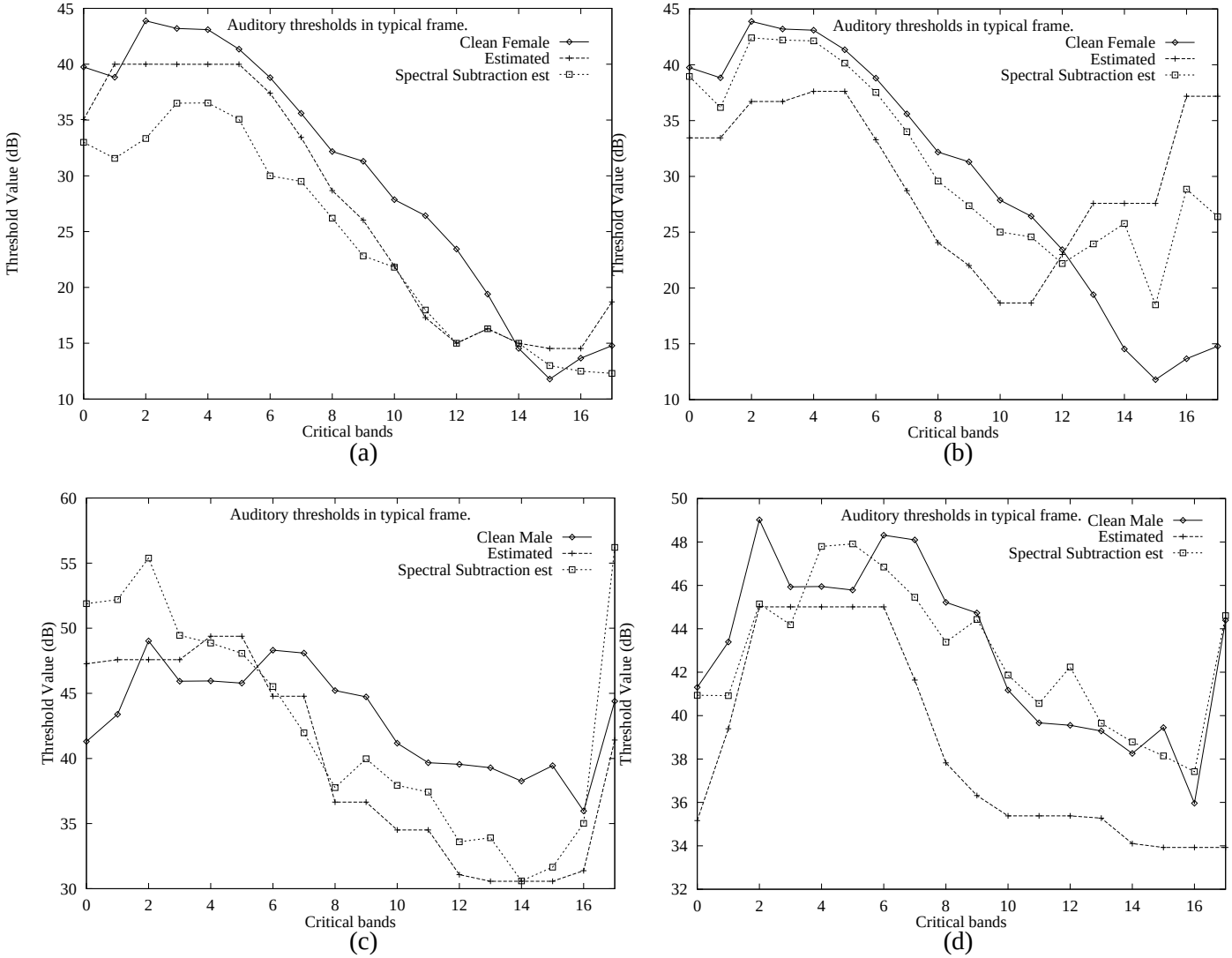


Figure E.5: *Estimate of thresholds for (a) Female corrupted by male speech (b) Female corrupted by* $1kHz$ *tone (c) Male corrupted by pink noise (d) Male corrupted by music all at* $5dB$

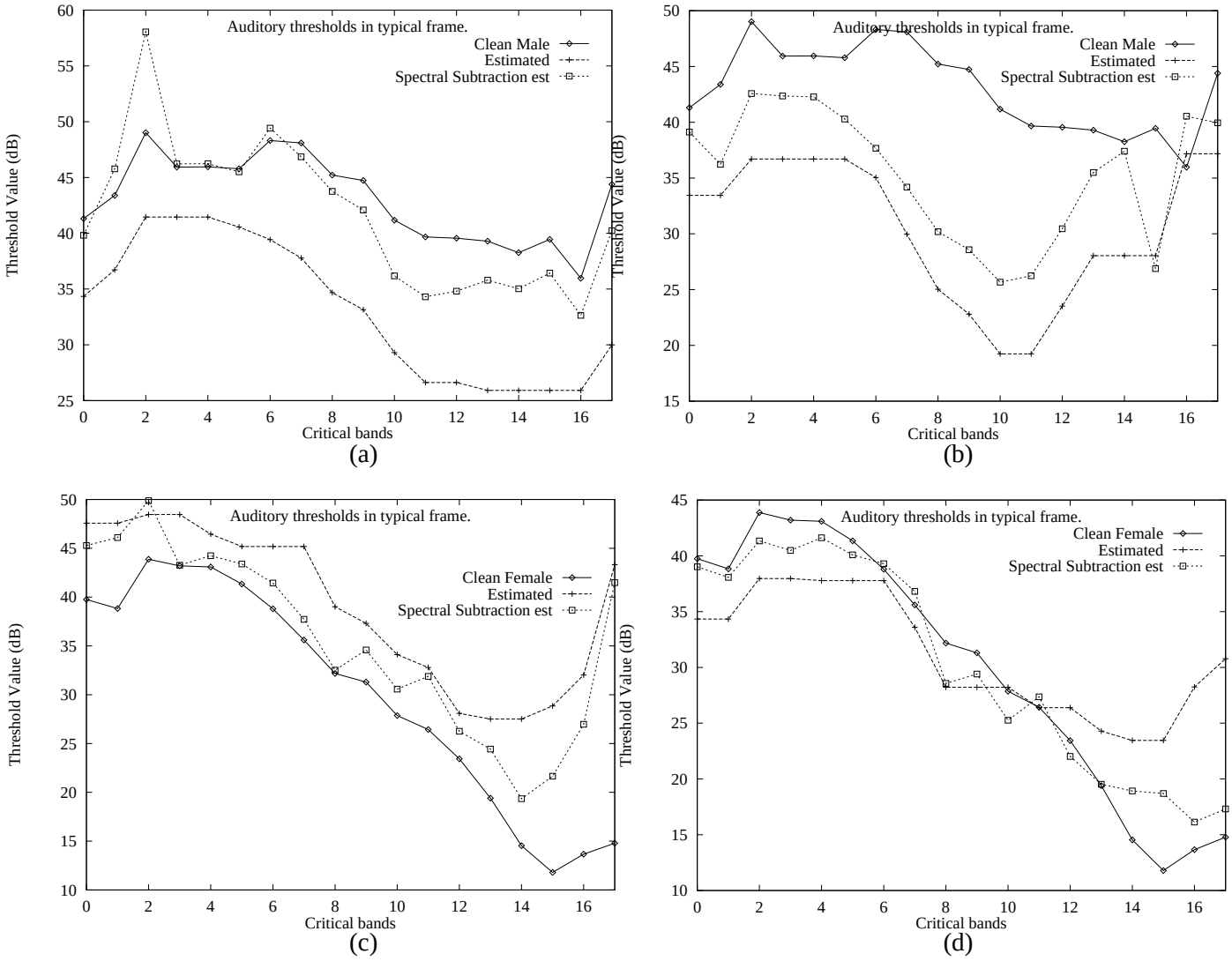


Figure E.6: *Estimate of thresholds for (a) Male corrupted by female speech (b) Male corrupted by* $1kHz$ *tone (c) Female corrupted by pink noise (d) Female corrupted by music all at* $5dB$

Appendix F
Robustness Noise Examples.

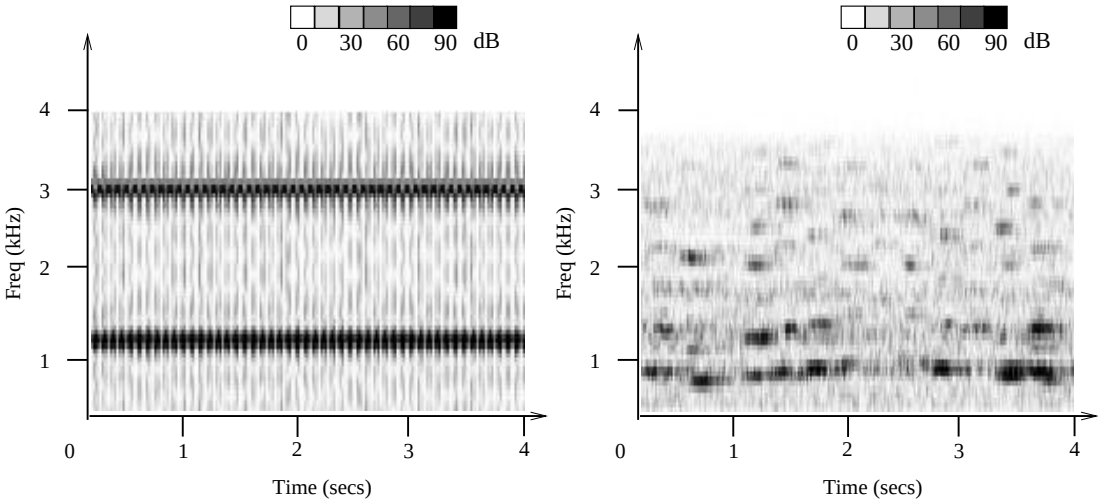


Figure F.1: *(a) Reference Noise test signal 1 (b) Reference Noise test signal 2*

Appendix G
Published Work.

- M A Tuffy and D I Laurenson, *"Background Noise reduction for Mobile Telephony"*, Published in the Proceedings of the 1998 IMA International Conference on Mathematics in Communications. December 1998.
- M A Tuffy and D I Laurenson, *"Estimating Clean Speech Masking Thresholds For Perceptual Based Speech Enhancement"*, Published in the Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp 127-130, October 1999.
- M A Tuffy and D I Laurenson, *"Speech Enhancement in a Non-Stationary Environment"*, submitted to the IEEE Transactions on Speech and Audio Processing.

Background noise reduction for Mobile Telephony

Mark A. Tuffy and D. I. Laurenson
Department of Electronics and Electrical Engineering,
The University Of Edinburgh, Edinburgh

Abstract

A new technique for speech enhancement is proposed which uses psychoacoustic analysis and multiple noise estimation techniques. This technique enhances the standard spectral subtraction method by the use of auditory masking thresholds, to determine spectral characteristics unaffected by corrupting background noise. The advantage of such a system is the ability to closely estimate rapidly changing noise characteristics in mobile phone conversations, for which this technique was developed.

1. Introduction

A great deal of research has been conducted into speech enhancement systems to remove background noise from hands-free cellular systems for cars [1,2], but little has been aimed at mobile users moving through an environment where there are changing noise conditions of various types. The potential for current techniques which can be applied to such a problem is diverse,

The algorithm presented here is based on the spectral subtraction method as reported by Boll [3]. This assumes that the noise is additive and uncorrelated to the speech. This also requires no *a priori* information on the corrupting noise, and requires only its level (as presented by a reference microphone), and its frequency distribution.

To improve both the SNR and intelligibility of the processed speech an algorithm must also take into account the perceptual characteristics of speech. This has been done in various ways [4-6] with degrees of success. In any system which integrates perceptual criteria into spectral subtraction there is the need to obtain *a good noise estimate, acceptable noise reduction, minimal Spectral distortion in the processed speech and reasonable computational time*.

2. Method Used

This technique is designed to be used as a preprocessor for a mobile phone. Single sensor noise reduction systems [7] are limited in application, thus a two sensor approach is adopted where the input to one sensor is the corrupted speech, and the other records the environmental noise at that time. The time domain additive noise corrupted speech is described as

$$s_n(t) = s(t) + n(t) \tag{1}$$

where $s_n(t)$ is the corrupted time domain speech, $s(t)$ is the clean speech signal, and $n(t)$ is the additive environmental noise.

As the spectral subtraction and perceptual techniques are frequency domain approaches operating only on the magnitude of the signal, it is transformed using the short-time discrete Fourier transform (STDFT). This is required as most audio signals do not remain statistically stationary for the period of measurement [8]. For speech the spectrum changes every few milliseconds meaning that long term measures are not appropriate. In terms of noise estimation it is also important to present as accurate a reference to the spectral subtraction as possible implying that short term analysis is more appropriate.

The experimental data were speech and noise signals 10 seconds long sampled at 8kHz. In general speech analysis is performed on frames of data no more than 20ms such that the speech can be assumed stationary within the frame, while the rapidly changing

noise remains non-stationary. This is an advantage over one sensor systems where the noise updates are only detected during speech pauses and assumed stationary between each update. In this case the STDFT was implemented on 128 sample Hamming multiplied windows with a 50% overlap. Thus in the frequency domain,

$$S_N(\omega) = \mathrm{STDFT}[s_n(t).w(t)] \tag{2}$$

where $S_N(\omega)$ is the frequency domain corrupted speech,and $w(t)$ is the time domain Hamming window.

2.1. Perceptual Modelling

From study of the human auditory process, it is apparent that humans have the ability to suppress concurrent sounds to pick out those which are of interest, the so called 'cocktail party effect'. While some of this suppression is due to neural processing, the auditory system produces masking thresholds which determine the spectral features which are audible.

Masking depends on the relative levels of signal and noise that are presented to the auditory system. From this the masking thresholds are derived and the audible spectral values can then be found. The model used in this algorithm is that presented by Johnston[9]. From Figure 1 it can be seen that noise estimation is a vital part of the speech

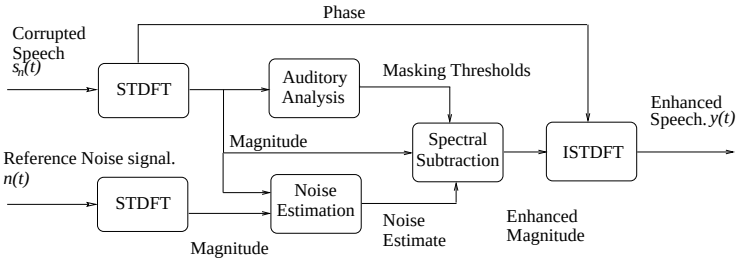


Figure 1. Noise suppression algorithm.

enhancement algorithm. In many of the speech enhancement systems developed today, there is a great emphasis placed on how the noise estimation is developed. One thing that is common to all techniques though is the concept of using one estimation algorithm.

One of the new noise estimation techniques presented here takes in frames of background noise, as well as the corrupted speech and passes it through 5 different estimation algorithms in parallel.

2.2. Noise estimation.

2.2.1. Techniques used

The 5 techniques were chosen for their approach to the noise estimation, along with the level of computational complexity they required. While the system is not ready to implement in a mobile phone, it was prudent to look at low level designs which may eventually lead to real time performance.

2.2.2. Boll's Noise Averaging Algorithm

In [3] Boll presents his well known noise averaging algorithm. In this the estimated noise magnitude $|N(\omega)|$ is determined by taking the average value of the reference noise $\mu(\omega)$, during pauses in the speech. These pauses were detected by the comparison of the energy content of the corrupted speech to $\mu(\omega)$.

2.2.3. Abdel, Mokhtar and Ezz-Al-Arad

In [10] an approach is presented to enhance speech in communication between vehicles. This is a short time spectral amplitude estimator

$$P_{\widehat{S}}(\omega) = P_{S_N}(\omega) - rP_N(\omega) \tag{3}$$

where $P_{\widehat{S}}(\omega), P_{S_N}(\omega), P_N(\omega)$ are the power spectra of the enhanced speech, the corrupted speech and the environmental noise respectively and r is the subtraction factor. [10] states that an overestimation of the noise *i.e.* $(r >> 1)$ would result in superior performance than the case $r = 1$. In this case the optimal value of r is taken to be 3 as in [10].

No speech or pause detection is implemented in this case, although one merit of this approach is the adoption of a spectral noise floor. As in [3] it is suggested that a small percentage of the noise replace negative values of $P_{\widehat{S}}(\omega)$. This reduces the occurrence of *"musical noise"* artifacts in the processed speech i.e.

$$P_{\widehat{S}}(\omega) = \begin{cases} P_{S_N}(\omega) - 3P_N(\omega) & : \quad P_{S_N}(\omega) \geq 3P_N(\omega) \\ \eta P_N(\omega) & : \quad P_{S_N}(\omega) < 3P_N(\omega) \end{cases}$$

where η is the noise spectral threshold, which is set to a value of $\eta = 0.008$ in [10].

2.2.4. Pollak, Sovka and Uhlir

In [2] an algorithm based on energy tracking is presented. The energy present in a window of corrupted speech and background noise is calculated as

$$E_{S_N} = \sum_{n=0}^{N-1} S_N^2(\omega), E_N = \sum_{n=0}^{N-1} N^2(\omega) \tag{4}$$

where $n = 0, ..., 128$, E_{S_N} = energy of the corrupted speech window and, E_N = energy of the noise window. From this an energy threshold, E_{th}, is calculated as $E_{th} = 1.5E_N$. If $E_{S_N} > E_{th}$ then speech is present, otherwise only noise is present. When a noise update is required, $E_{ref}^{new} = (1-p)E_{ref}^{old} + pE_{s_N}$ where p is the forgetting factor and E_{ref}^{new},E_{ref}^{old}, are the new and previous noise estimates respectively. The value of p is altered to minimise the musical artifacts in the speech.

2.2.5. Kang and Fransen

Kang and Fransen propose that it is only the signal energy below 1kHz that is required to calculate speech activity [11]. As each new frame is analysed the past history of all frame lowband (i.e.below 1kHz) energies are scanned to determine MAX and MIN energy values. When the current energy is below a certain threshold the noise spectrum is updated.

$$Thresh = MIN + \frac{MAX - MIN}{8} \tag{5}$$

$$P_N^{new}(\omega) = GP_N^{old}(\omega) + (1-G)P_{S_N} \tag{6}$$

where G is the feedback factor which is normally $\frac{15}{16}$

2.2.6. Noise estimation with a hangover mechanism (NEAH)

It was decided to introduce a new algorithm developed very early in this work on spectral subtraction. As with many of those shown here it uses noise estimation for the speech or pause detection, and a comparison is made between the frames of noise and speech.

One concern is the possibility that frames of speech could be incorrectly detected as noise, due to the corruption caused. To prevent this case, if a frame is detected as noise, the next frame is then examined before any changes are made. If the next frame is also noise then this was a candidate for update and not a product of speech corruption. This is implemented as follows;

$$\text{if} \quad E_{S_N} < E_N \qquad \text{Wait for next frame.}$$
$$\text{if} \quad E_{S_N}^{next} < E_N^{next} \qquad \text{Noise so update.}$$

When the update occurs the noisy spectrum magnitude is taken as the noise estimate. All other times it is the background estimate.

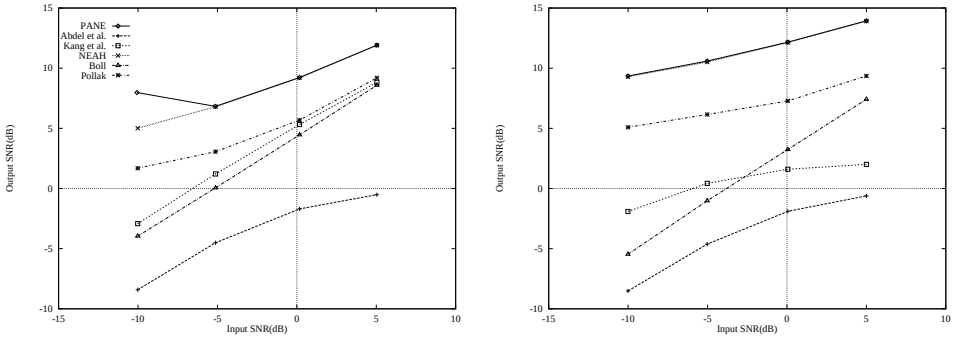
2.3. Parallel Noise estimation

While each individual algorithm may work well in a specified context, the developed technique is required to work in a number of conditions. To determine if better performance could be achieved through a combination of the techniques shown here, an algorithm was written which runs all five in parallel and a maximum likelihood case is applied to see if they agree that a frame of noise is present. When the majority agree an update to the estimate is performed by taking the maximum of the respective frame values.

3. Results

To test the performance of the techniques female speech was mixed with pink noise or a competing male speaker at four SNRs (-10dB, -5dB, 0dB, 5dB). Pink noise is used as it is considered to be a good estimate for the conditions present in general environmental noise. The speech enhancement routine is tested using each estimate technique on its own, and then all in combination for the parallel case. Figures 2(a) and (b) show the SNR results for the pink noise and competing speaker cases, Figures 4(a) and (b) show the rms magnitude error results which are an indication of the distortion introduced by the algorithm. It should be noted that in order to compare algorithms with similar complexities the results presented in Figures 2 (a) and (b) deal with magnitude spectral subtraction whilst the techniques suggested by Kang *et al.* and Abdel *et el.* were originally implemented in an algorithm using power spectral subtraction. Figure 3 shows the comparative performance of these algorithms operating in the way they were originally conceived.

In Figure 2(a) it can be seen that as expected the output SNR improves for each individual noise estimation technique as the input SNR increases for the pink noise corruption.



(a) (b)

Figure 2. (a) Pink Noise and (b) Competing speaker SNR results.

The similar performances of the Kang *et al.* and Pollak *et al.* approaches is not unexpected, as they approach the estimation in a similar way. They only differ in the

very low SNR regions leading to the belief that Kang's approach is not as useful in this case. Boll's Algorithm is not far behind these two in terms of SNR improvement which is a surprise given that it is a simple noise averaging approach. The two new algorithms presented here though outperform any of the other algorithms, with at least a 3 to 4 dB improvement over the new energy technique, and as much as 5dB more than the next nearest system. What is noticeable though is a drop in the results for the parallel case between -10dB and -5db input SNR. This is thought to be due to the differing error sources of the techniques used in the parallel combination, which could lead to a diversity gain at a lower SNR. This drops as the input SNR increases and is backed up by the error results presented in Figures 4(a) and (b).

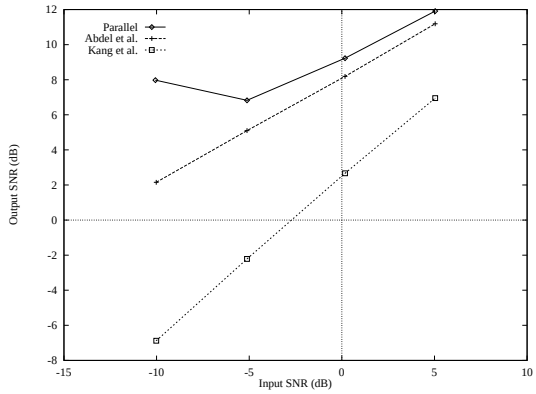
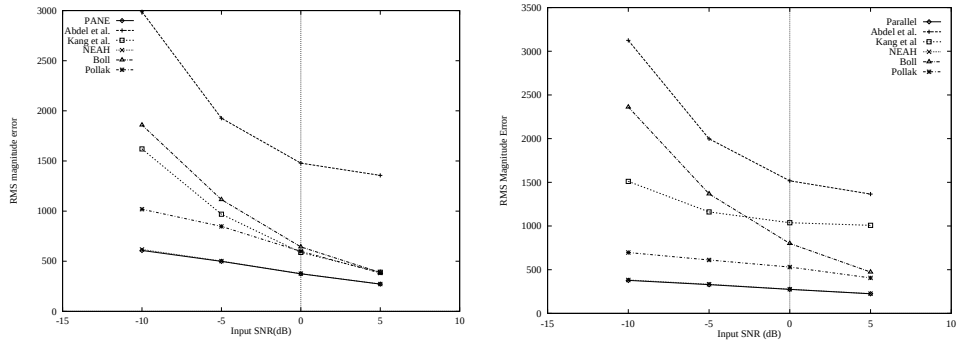


Figure 3. Pink noise power spectral subtraction SNR results for Abdel and Kang et al.

Figure 3 shows the performance of the parallel algorithm as calculated above along with that of Kang *et al.* and Abdel *et al.* in a power spectral subtraction experiment. It must be noted thought that the plot of the parallel case still refers to magnitude spectral subtraction and is presented for comparison. Comparing the results of Kang *et al.* to the magnitude case and there seems to be little improvement, indeed at the lower SNR the performance is reduced. The largest difference though can be seen in the improvement of the Abdel *et al.* algorithm which gains up to 11dB through the use of power spectral subtraction. Despite this improvement the parallel combination using only magnitude spectral subtraction still outperforms these techniques in this experiment.

For the competing speaker case shown in Figure 2(b) again, the results show the new parallel and NEAH algorithms perform better than any of the rest. In this case though it is worth noting that unlike the pink noise case the two techniques show almost identical results. It is also shown that the approach by Kang *et al.* performs significantly worse in this test than in the pink noise case. The rms magnitude error graphs (Figures 4(a) and (b)) are presented to give an indication of the quality of the enhanced speech. No formal subjective tests have been carried out at this time, however these are part of future work planned to improve the performance.

As expected the algorithm presented by Abdel *et al.* produces the largest rms errors, although the results improve as the input SNR rose. The rms results confirm the conclusions drawn from the SNR case with the new algorithms performing better than any of the techniques studied. Figures 4(a) and (b) show this for both the pink noise and competing speaker scenarios, with degradation in the Kang *et al.* approach also shown for the rms magnitude error. The increased performance of Abdel *et al.* in terms of SNR for power



(a) (b)

Figure 4. (a) Pink noise and (b) Competing speaker rms magnitude error results.

spectral subtraction is negated by the fact that this algorithm produced the worst rms errors. This proves the point that in terms of speech assessment an increase in SNR alone does not imply that better quality speech has been produced at the output.

4. Conclusion

Two new noise estimation algorithms have been produced, an improved energy technique (NEAH) and a parallel noise estimation technique. Both of these have been shown to give superior performance to some established techniques, in terms of both SNR and rms magnitude error for pink noise or competing speaker corruption. The performance of NEAH can give up to 3dB improvement over the nearest established algorithm, while the parallel approach gives up to 5dB improvement. Both of these new algorithms also introduce considerably less spectral distortion than any of the others, leading to the conclusion that objectively speech quality is maintained better than any of the other algorithms presented.

References

[1] S.M.Kuo, J.Kundruru, *Sub-band Adaptive Noise Cancelling for Hands-Free Cellular Phone Applications*, 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.19-22.

[2] P.Pollak, P.Sovka, J.Uhlir, *Noise Suppression system for a Car*, EUROSPEECH'93, pp.1073-1076, Sep 1993.

[3] S.F.Boll, *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Trans ASSP, ASSP-27, pp.113-120, April 1979.

[4] Y.M.Cheng, D.O'Shaughnessy *Speech Enhancement Based Conceptually on Auditory Evidence*, IEEE Trans on Signal Processing, Vol 39, No 9, pp.1943-1954, Sept 1991.

[5] T.Usagawa, M.Iwata, M.Ebata, *Speech Parameter Extraction in Noisy Environment Using a Masking Model*, Proc. ICASSP'94, Vol 2, pp.81-84.

[6] D.Tsoukalas, M.Paraskevas, J.Mourjopoulous, *Speech Enhancement Using Psychoacoustic Criteria*, Proc. ICASSP'93, Vol2, pp.359-366.

[7] E.B.George, *Single Sensor Speech Attenuation using a Soft Decision/Variable Attenuation Algorithm*, Proc ICASSP'95, Vol 1, pp.816-819.

[8] J.R.Deller, J.G.Proakis, J.H.L.Hansen, *Discrete-Time Processing of Speech Signals*, MacMillen Publishing, 1993, ISBN 0-02-328301-7.

[9] J.D.Johnston, *Transform Coding of Audio Signals Using Perceptual Noise Criteria*, IEEE Journal on Selected Areas of Communications, Vol 6, No 2, pp.314-323, Feb 1988.

[10] A.O.Abdel, M.A.Mokhtar, M.A.Ezz-Al-Arab, *Speech Enhancement in the communication between vehicles*, 39th IEEE Vehicular Technology Conference, Vol 2, pp.897-901, May 1989.

[11] G.S.Kang, l.J.Fransen, *Quality Improvement of LPC Processed Noisy Speech by Spectral Subtraction*, IEEE Trans ASSP, Vol ASSP-37, pp.939-942, June 1989.

Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

ESTIMATING CLEAN SPEECH THRESHOLDS FOR PERCEPTUAL BASED SPEECH ENHANCEMENT.

Mark A. Tuffy

Signals and Systems Group
Dept. Of Electronics and Electrical Engineering
The University Of Edinburgh
King's Buildings, Mayfield Road
Edinburgh EH9 3JL, Scotland.
mat@ee.ed.ac.uk

D. I. Laurenson

Signals and Systems Group
Dept. of Electronics and Electrical Engineering
The University Of Edinburgh
King's Buildings, Mayfield Road
Edinburgh EH9 3JL, Scotland.
Dave.Laurenson@ee.ed.ac.uk

ABSTRACT

In this paper we propose a new method for the estimation of clean speech masking thresholds for speech enhancement. These thresholds are applied to a perceptually based spectral subtraction algorithm to enhance speech in a non-stationary noise environment. In contrast to other approaches we do not directly use an estimate of the clean speech to obtain the masking thresholds, but examine the relationship between those from the corrupted speech and corrupting noise.

The thresholds from this algorithm are compared to those produced from a clean speech estimate from a variety of common spectral subtraction algorithms.

1. INTRODUCTION

Since S.F.Boll[1] first introduced spectral subtraction in the late 70's, it has been the basis for many speech enhancement algorithms. Until recently though little work had been done to expand this approach, especially with a view to make the output perceptually acceptable. There has recently been an increase in work to harness auditory analysis, either for coding requirements [2] or for speech enhancement[3], and this has led to the introduction of perceptual criteria into spectral subtraction.

The key to this work is the application of auditory masking thresholds and their effect on which portions of a signal are audible or inaudible. Some of the best work in this area has been by J.D.Johnston [2] whose work is often found to be the basis of many perceptual systems in use today.

The advantages of applying auditory masking thresholds to any form of audio signal processing is obvious. In applications such as Mini-Disc thresholds allow the codec to determine which portions of a musical signal are audible and which are not. Those tagged as being inaudible (under the threshold) do not require to be coded, and this is where part of the data compression is achieved. In speech enhancement masking thresholds allow an increase in algorithm speed by operating only on audible points and takes into account the way the ear hears. In noise shaping it manipulates any residual noise into a more perceptually acceptable form.

In speech enhancement perceptual techniques encounter problems as there is normally no clean speech reference from which the thresholds can be calculated. To compound this there is the need for a good spectral estimate of the noise to ensure that the performance of the algorithm is maximised. The better the estimate, the lower the chance of speech distortion from the spectral subtraction. In a stationary noise environment this can be readily done, but the more non-stationary the noise the harder this becomes.

In this paper we discuss a new approach for estimating the clean speech thresholds required for speech enhancement. In contrast to other approaches the proposed one does not require any modification of the corrupted speech magnitude to calculate the thresholds. Instead the noise and the corrupted speech signal are used to calculate a pair of thresholds whose relationship is studied to determine the clean speech case.

2. MASKING THRESHOLDS CALCULATION.

For perceptually based spectral subtraction algorithms one of the most important aspects is an accurate estimate of the clean speech masking threshold. As can be seen in Fig 1, if the estimate is inaccurate the enhanced speech may either contain a high degree of residual noise, or cause a loss of speech information.

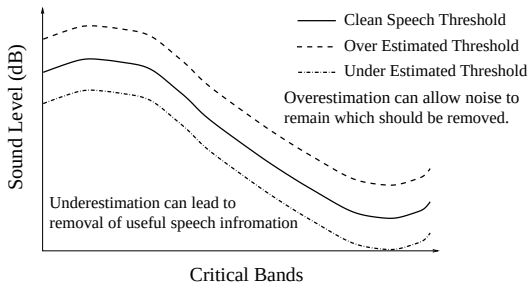


Figure 1: *Example of Clean Speech Threshold Estimation.*

In any condition where the user has the original clean speech this can of course be easily done. In real life situations though this is unlikely and an estimate of the clean speech must be produced. In general the technique of spectral subtraction is used to obtain a clean speech estimate as in [3]. This is then used to calculate the masking thresholds.

2.1. Spectral Subtraction

Given a noisy speech signal $y(n)$;

Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

$$y(n) = s(n) + n(n) \quad (1)$$

where $s(n)$ is the noisy speech signal and $n(n)$ the corrupting noise, spectral subtraction works on the magnitude of this resultant signal $Y(\omega)$ as obtained though a Short-Time Fourier Transform. The idea is to obtain as close an estimate of the short-time magnitude of the clean speech as possible. The generalised form of spectral subtraction is described as;

$$|\widehat{S}(\omega)| = (|Y(\omega)|^{\alpha} - \beta|\widehat{N}(\omega)|^{\alpha})^{1/\alpha} \quad (2)$$

where $Y(\omega), \widehat{S}(\omega), \widehat{N}(\omega)$are the magnitude spectra of the noisy speech, estimated clean speech and estimated noise respectively.

The parameters α and β can be chosen to tailor the algorithm for specific needs. When $\alpha = 2$ you have power spectral subtraction and with $\alpha = 1$ this becomes magnitude spectral subtraction. The oversubtraction factor β is normally chosen to try to prevent the introduction of musical noise into $\widehat{S}(\omega)$. The examination of which α and β values is a topic on its own right leading to work such as [4].

Having obtained the estimated clean speech signal there is the choice of what to do for those values which fall below zero as negative spectral values are not allowed. Techniques which are popular include full wave rectification, or the introduction of a noise floor. The latter is used in an attempt to perceptually whiten out any residual noise and make it less obtrusive to the listener. This is implemented to give the conditions

$$\widehat{S}(\omega) = |Y(\omega) - \beta\widehat{N}(\omega)|$$

for full wave rectification and

$$\widehat{S}(\omega) = \begin{cases} Y(\omega) - \beta(\widehat{N}(\omega)) & : \quad Y(\omega) \geq \eta\widehat{N}(\omega) \\ \eta P_N(\omega) & : \quad Y(\omega) < \eta\widehat{N}(\omega) \end{cases}$$

for the noise floor approach.

2.2. Masking Threshold Calculation

The approach most used for the masking threshold calculations is based on the work done by J.D.Johnston[2] and this has 4 sections;

- Critical band analysis
- Convolution with a spreading function
- Subtraction of the threshold offset
- Renormalisation and absolute threshold comparison

In speech enhancement the masking thresholds are used to determine those frequencies which do not require noise suppression as they are either masked by speech or inaudible. As shown in Fig 1 an inaccurate calculation of these could be more detrimental to the speech enhancement that improve the result.

3. NEW THRESHOLD ESTIMATION TECHNIQUE

Having looked at the methods used to calculate masking thresholds for speech enhancement one thing stands out. If the clean speech estimate is not accurate, then the masking thresholds derived from it could never be. This is especially true in adverse noise conditions where spectral subtraction is prone to producing many musical noise artifacts. These can fall into two main categories which are equally damaging for the calculation of masking thresholds.

Firstly, the random removal nature of spectral subtraction can result in large energy peaks in the magnitude spectra. This presents the problem of artificially raising the critical band energy (and hence masking threshold), associated with the affected frequency range. This results in the masking thresholds being set too high, and during enhancement noise that should be removed is erroneously tagged as being inaudible.

Secondly, there is the condition that half wave rectification or noise floor substitution can present. If a number of the points in the spectra are found to be below zero, half wave rectification will change these to zero resulting in a spectral null. The same effect although not as pronounced occurs if a spectral noise floor is substituted for these values. This has the effect of reducing the masking threshold (or in the case of half wave rectification producing errors) which can cause the removal of useful speech information.

It became clear that the use of an estimated clean speech magnitude $\widehat{S}(\omega)$ was not the optimum way to produce the thresholds. A new method was required which could sidestep the spectral subtraction produced thresholds. A possible solution to this is shown in Fig 2.

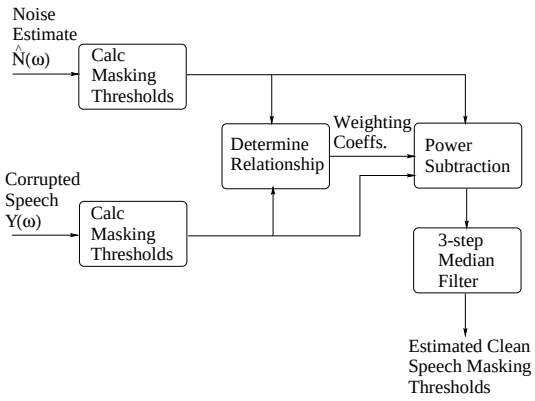


Figure 2: *Block diagram of the proposed method for estimating clean speech masking thresholds.*

The proposed algorithm calculates the masking thresholds produced by the corrupted speech and the reference noise. The relationship between these is examined and a weighting produced for each critical band in the window, which is then used in energy subtraction of the two thresholds. The advantage of such an approach is the inability to introduce errors from spectral subtraction into the threshold calculation. All the signals are manipulated in the non-linear auditory based critical band domain.

3.1. Calculation of the Weighting Factor

The focal point of the work presented here is the examination of the relationships between the thresholds produced from the corrupted speech and background noise. An example of such is shown in Fig 3.

It can be seen that the interaction between the thresholds for each critical band (CB) falls into the general categories;

- $SNR = 0dB$
- $SNR < 0dB$
- $SNR > 0dB$

Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

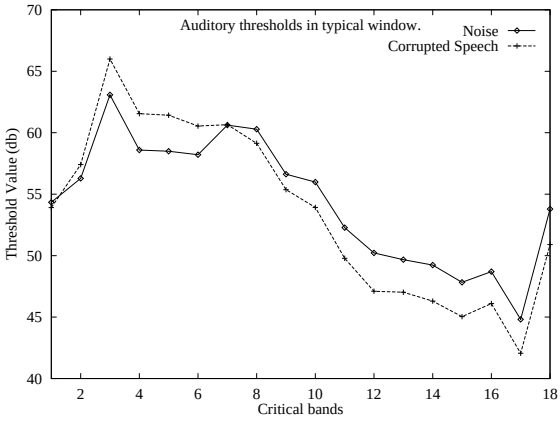


Figure 3: *Example of the relationship between the masking thresholds.*

Upon experimentation it was found that these general categories were not comprehensive enough to cover each CB interaction. It was found that the points around $0dB$ required different weightings from those with a SNR $> 3dB$ or $< -3dB$. This lead to a more specific categorisation;

1. $SNR = 0dB$
2. $-1 \leq SNR \leq 1dB$
3. $1 < |SNR| \leq 3dB$
4. $|SNR| > 3$

From these differing conditions weightings for the equation:

$$(|\widehat{E}_s(\omega)|) = E_{cs}(\omega) - \alpha E_n(\omega) \quad (3)$$

where $|\widehat{E}_s(\omega)|$, $E_{cs}(\omega)$, $E_n(\omega)$ and α are the estimated clean speech energy, the corrupted speech energy, noise energy and weighting factor respectively. After conversion back into decibels this gives the estimated clean speech thresholds.

3.1.1. $SNR = 0dB$

In this case the signal presented to the noise and speech microphone is the same. In this case the weighting applied to the is taken as one meaning a full energy subtraction. This case is only likely in those windows where there is no speech present and to counter the possibility of a threshold null in these cases the calculated threshold is compared to the Minimum Audible Pressure for that CB (MAP) as defined in [5] and if below the MAP is substituted for the derived value.

3.1.2. $-1 \leq SNR \leq 1dB$

For all the CB's in this range (not including the $0dB$ cases as described above) there is an obvious speech masking threshold component present due to the fact that the $SNR \neq 0dB$. In this case it was felt that the values of the corrupted speech threshold should be slightly emphasised, as once the majority of the noise was removed these would closely track the clean speech case. To achieve this α was set to be 0.8.

3.2. $1 < |SNR| \leq 3dB$

When the outer limits of this case are reached one of the critical bands holds twice as much signal power as the other. It was found that many of the points in the comparison can be clustered in this case as can be seen in Fig 3. Rather than setting one α for this wide range of points a soft allocation of α was performed. This was calculated by normalising the energy ratio of the critical bands taking 2 ($3dB$) or 0.5 ($-3dB$) to be the maximum. Again in an attempt to emphasise the contribution of the corrupted speech thresholds a value of half of this normalised factor is taken to be α.

$$E_{nor}(\omega) = \frac{E_{ratio}(\omega)}{\beta} \quad (4)$$

$$\alpha = 0.5(E_{nor}(\omega)) \quad (5)$$

where $E_{nor}(\omega)$, $E_{ratio}(\omega)$, β are the normalised energy ration, the energy ratio and the maximum limit (either 0.5 or 2) respectively.

3.2.1. $|SNR| > 3$

In the case where the SNR falls into this category then there is a considerable difference in the thresholds due to the influence of the speech present in the CB. Whether the SNR is negative or positive once it has reached this stage it is better to follow in some way the pattern of the corrupted speech derived thresholds. While some energy from the noise may be present the general shape of the thresholds will be a good approximation to the clean speech case. In such cases the energy of the corrupted speech threshold is dropped by a factor γ. Through experimentation it has been found that for $SNR > 3dB$ a good value for $\gamma = 4$ and for $SNR < -3dB$ $\gamma = 3$

4. RESULTS

The algorithm was tested with male and female speech recorded at 8kHz and the corrupting noise consisted of pink noise, music, a competing speaker of opposite sex and a 1kHz tone. The SNR's varied from $-15dB$ to $10dB$. A selection of graphical results are shown here as this is the best way to show the performance, and a comparison is made with the power spectral subtraction method utilising a noise floor factor of $\eta = 0.08$. It is interesting to note is that in Figs 4 and 5 the algorithm has been able to make a good estimate of the clean female speech, and is slightly more conservative than the standard power spectral subtraction based method. In the case of the result show in Fig 4, less uncorrupted speech information is presented to the enhancement process as erroneously detected noisy speech points.

In Fig 5 the conservative nature of the algorithm means that some noisy speech points are passed unaltered through the algorithm. This is acceptable as a small amount of noise in any resultant enhanced speech will help to reduce the possibility of musical noise. In Figs 6 and 7 though it can be seen that the algorithm has a harder time estimating the thresholds for male speech. In Fig 6 it can be seen in CB 3-6 the estimate is too high and the algorithm also finds it hard to track these CB's in Fig 7. In Fig 6 the only portion of the female speech that was not predicted as well was in CB's 15-18, which is the higher frequecy region. What is marked about comparison between these pairs of results is the ability of the algorithm to estimate female thresholds far more accurately than male thresholds. Further work is being done to determine if this is a parallel of the results described by [6] into the effects the type

Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

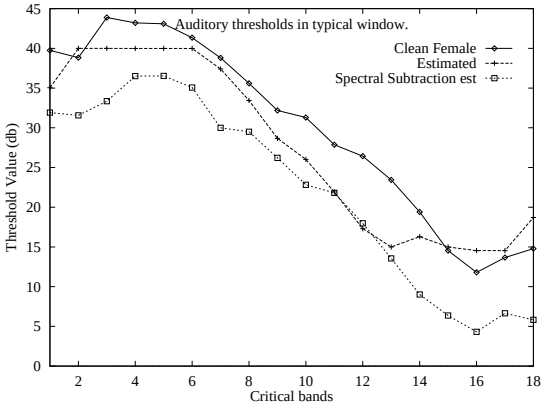


Figure 4: *Estimated Threshold for Clean Female Speech after corruption by Male speech SNR* $5dB$.

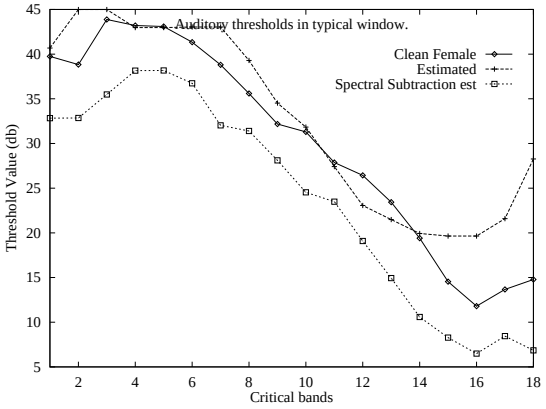


Figure 5: *Estimated Threshold for Clean Female Speech after corruption by Male speech SNR* $0dB$.

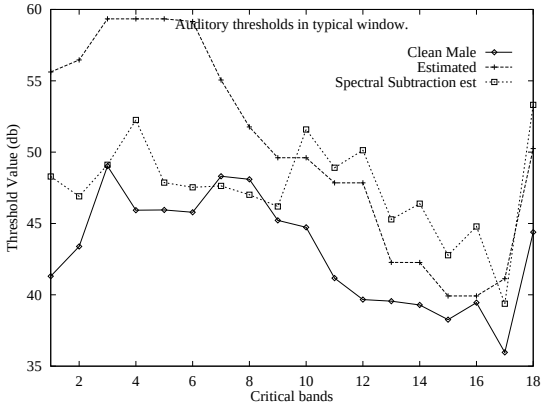


Figure 6: *Estimated Threshold for Clean Male Speech after corruption by Pink Noise SNR* $-10dB$.

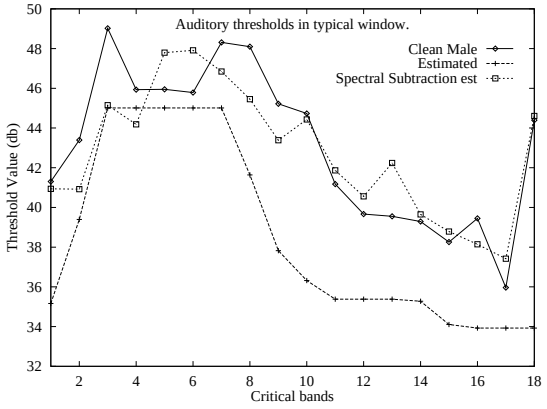


Figure 7: *Estimated Threshold for Clean Male Speech after corruption by music SNR* $0dB$.

of noise and sex of speaker has on the intelligibility of speech by humans. The quality of the speech produced by the algorithm compares favourably with that produced with the spectral subtraction approach. The overestimation of the higher frequqency CB's for the female speech leads to a slight sharpening of the output. The male speech generally suffers from a increase in the background noise artifacts left in the speech, although no speech information has been removed.

5. CONCLUSIONS

A new method for the estimation of clean speech masking thresholds for speech enhancement has been developed. It has the advantage of not operating on an estimate of clean speech, hence preventing the effects of musical noise or spectral nulls on the threhsold calculation. Further work is being done to examine the differences between the performance for male and female speakers, although the algorithm can be seen as a valid alternative to the established spectral subtraction based one.

6. REFERENCES

[1] S.F.Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. ASSP, ASSP-27, pp.113-120, April 1979.

[2] J.D.Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria", IEEE Journal on Selected Areas of Communications, Vol 6, No 2, pp.314-323, Feb 1988.

[3] D.Touskalas, M.Parashevas, J. Moujopoulous, "Speech Enhancement Using Perceptual Criteria", Proc. ICASSP '93, Vol 2, pp.359-362, April 1993.

[4] R.D.Niederjohn, P. Lee, F.Josse, "Factors Related To Spectral Subtraction for Speech in Noise Enhancement", Proc. IEEE International Conference on Industrial Electronics and Instrumentation pp.985-996, 1987.

[5] M.C.Killion, "Revisied estimate of minimum audible pressure: Where is the "missing 6 dB"?", Journal of the Acoustical Society of America, Vol 63(5), pp.1501-1508, May 1978.

[6] J-C Junqua, "The Influence of Psychoacoustic and Psycholinguistic Factor On Listener Judgements of Intelligibility of Normal and Lombard Speech.", Proc ICASSP 91, Vol1, pp 361-364, 1991.

References

[1] E. Lindemann, "Two microphone nonlinear frequency domain beamformer for hearing aid noise reduction," in *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 24–27, 1995.

[2] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*. New York: Marcel Dekker Publishing, 1998, ISBN 0-8247-0143-7.

[3] W. A. Yost, *Fundamentals of Hearing: An Introduction, 3rd Edition*. London: Academic Press, 1994, ISBN 0-12-772690.

[4] S.M.Kuo and J.Kundruru, "Sub-band adaptive noise cancelling for hands-free cellular phone applications," *1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 19–22, 1993.

[5] S.Nordebo, S.Nordholm, B.Bengtsson, and I.Claesson, "Noise reduction using an adaptive microphone array in car-a-speech recognition evaluation," *1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 16–18, 1993.

[6] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[7] B. Widrow, J. R. G. Jr, J. M. McCool, J. Kaunitz, R. H. H. C S Williams, J. R. Zeidler, E. D. Jr, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *IEEE Proceedings*, vol. 63, pp. 1692–1717, December 1975.

[8] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete Time Processing of Speech Signals*. New York: Macmillan Publishing Company, 1993, ISBN 0-02-328301-7.

[9] J. J. Rodriguez and J. S. Lim, "Adaptive noise reduction in aircraft communication systems.," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 169–172, 1987.

[10] W. A. Harrison, J. S. Lim, and E. Singer, "A new application of adaptive noise cancellation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 43, pp. 21–27, February 1986.

[11] G. A. Powell, P. Darlington, and P. D. Wheller, "Practical adaptive noise reduction in aircraft cockpit environment.," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 172–176, 1987.

[12] R. Martin and J. Altenhoner, "Coupled adaptive filters for acoustic echo control and noise reduction," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 3043–3046, 1995.

[13] P. Dreiseitel, E. Hansler, and H. Puder, "Acoustic echo and noise control- a long lasting challenge," in *Proceedings of EUSIPCO*, pp. 945–952, 1998.

[14] C. Beaugeant and P. Scalart, "Combined systems for noise reduction and echo cancellation," in *Proceedings of EUSIPCO*, pp. 957–960, 1998.

[15] S. Gustafsson and P. Jax, "Combined residual echo and noise reduction: A novel psychoacoustically motivated algorithm.," in *Proceedings of EUSIPCO*, vol. 2, pp. 961–964, 1998.

[16] M. W. Hoffman, "Robust adaptive processing of microphone array data for hearing aids," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 1993.

[17] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics Speech and Signal Processing Magazine*, pp. 4–24, April 1988.

[18] T. Chen, "The past, present and future of audio signal processing," *IEEE Signal Processing Magazine*, pp. 30–57, September 1997.

[19] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal processing technique to remove room reverberation from speech signals," *Journal of the Acoustical Society of America*, vol. 62, pp. 912–915, October 1977.

[20] P. Yanick and H. Drucker, "Signal processing to improve intelligibility in the presence of noise for persons with a ski slope hearing impediment," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 24, pp. 507–512, December 1976.

[21] J. J. Rosowski, "The effects of external and middle ear filtering on auditory threshold and noise induced hearing loss," *Journal of the Acoustical Society of America*, vol. 90, pp. 124–135, July 1991.

[22] O. M. M. Mitchell, C. A. Ross, and G. H. Yates, "Signal processing for a cocktail party effect.," *Journal of the Acoustical Society of America*, vol. 50, pp. 656–660, May 1971.

[23] K. Farrell, R. J. Mammone, and J. L. Flanangan, "Beamforming microphone arrays for speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 285–288, March 1992.

[24] R. Zelinski, "Noise reduction based on microphone array with lms adaptive post filtering," *Electronic Letters*, vol. 26, pp. 2036–2037, November 1990.

[25] F. Ehrmann, L. Bouquin-Jeannes, and G. Faucon, "Optimisation of a two sensor noise reduction technique," *IEEE Signal Processing Letters*, vol. 2, pp. 108–110, June 1995.

[26] R. L. Bouquin and G. Faucon, "Using the coherence function for noise reduction," *IEEE Proceedings*, vol. 139, pp. 276–280, June 1992.

[27] D. Banks, "Localisation and separation of simultaneous voices with two microphones," *IEEE Proceedings*, vol. 140, pp. 229–234, August 1993.

[28] I. Aleksander and M. J. D. Wilson, "Adaptive windows for image processing," *IEEE Proceedings*, vol. 132, pp. 233–245, September 1985.

[29] J. C. Russ, *The image processing handbook.* CRC Press, 1995, ISBN 0849325161.

[30] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 30, pp. 679–681, August 1982.

[31] H. Hermansky, E. A. Wan, and C. Avendano, "Noise suppression in cellular communications," in *Proceeding of the 2nd IEEE IVTTA workshop*, pp. 85–88, September 1994.

[32] H. Hermansky, E. A. Wan, and C. Avendano, "Speech enhancement based on temporal processing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 405–408, March 1995.

[33] H. Hermansky, N. Morgan, and H. G. Hirsch, "Recognition of speech in additive and convolutional noise based on rasta spectral processing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 83–86, March 1993.

[34] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. New Jersey, US: Prentice-Hall, 1978, ISBN 0-13-213603-1.

[35] P. Pollák and P. Sovka, "The study of speech/pause detectors for speech enhancement methods.," in *Proceedings of the 4th European Conference on Speech Communication and Technology*, pp. 1575–1578, September 1995.

[36] P. Pollák, P. Sovka, and J. Uhlíř, "Cepstral speech/pause detectors.," in *Proceedings of the IEEE Workshop on Nonlinear Signal and Image Processing*, pp. 388–391, June 1995.

[37] S. V. Vaseghi, B. P. Milner, and J. J. Humphries, "Noisy speech recognition using cepstral-time features and spectral time filters," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 65–68, 1994.

[38] J. P. Openshaw and J. S. Mason, "On the limits of cepstral features in noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 49–52, 1994.

[39] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise.," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 26, pp. 471–472, October 1978.

[40] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *IEEE Proceedings*, vol. 67, pp. 1586–1604, December 1979.

[41] R. J. Niederjohn, P. J. Lee, and F. Josse, "Factors related to spectral subtraction for speech in noise enhancement," in *Proceedings of the IEEE International Conference on Industrial Electronics and Instrumentation*, pp. 985–996, 1987.

[42] O. Cappè and J. Laroche, "Evaluation of short time spectral attenuation techniques for the restoration of musical recordings," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 84–93, January 1995.

[43] O. Cappè and J. Laroche, "Enhancing speech degraded by additive noise or competing speakers," *IEEE Communications Magazine*, vol. 27, pp. 40–52, February 1989.

[44] S. M. McOlash, R. J. Niederjohn, and J. A. Heinen, "A spectral subtraction method for the enhancement of speech corrupted by nonwhite, nonstationary noise.," in *21st International conference on Industrial Electronics, Control, and Instrumentation*, vol. 2, pp. 872–877, 1995.

[45] B. Porat, *A Course in Digital Signal Processing*. Wiley Press, 1997, ISBN 0471149616.

[46] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, pp. 14–38, October 1991.

[47] G. Strang, "Wavelets appendix 1," *American Scientist*, vol. 82, pp. 250–255, April 1994.

[48] A. Graps, "An introduction to wavelets," *IEEE Computational Science and Engineering*, vol. 2, Summer 1995.

[49] D. E. Tsoukalas, J. N. Mourjopolous, and G. Kokkinakis, "Perceptual filter for audio signal enhancement," *Journal of the Acoustical Society of America*, vol. 45, pp. 22–36, January/February 1997.

[50] K.Gosse, F. M. de Saint-Martin, X. Durot, P. Duhamel, and J. B. Rault, "Subband audio coding with synthesis filter minimising a perceptual distortion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 347–350, 1997.

[51] J. D. Johnston and K. Brandenburg, *Wideband Coding- Perceptual Considerations for Speech and Music, ch4 in Advances in Speech Signal Processing*. Dekker Press, 1992, ISBN 0824785401.

[52] B. Espinoza-Varas and S. V. Cherukuri, "Evaluating a model of auditory masking for applications in audio coding," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 195–197, 1995.

[53] Y. Mahieux, "High quality audio transform coding at 64 kbits/s," *Ann. Télécommun*, vol. 47, no. 3-4, pp. 95–109, 1992.

[54] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 115–132, January 1994.

[55] C. Colomes, M. Lever, J. B. Rault, and Y. F. Dehery, "A perceptual model applied to audio bit-rate reduction," *Journal of the Acoustical Society of America*, vol. 43, pp. 223–240, April 1995.

[56] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *IEEE Proceedings*, vol. 81, pp. 1385–1422, October 1993.

[57] D. Sen, D. H. Irving, and W. H. Holmes, "Use of an auditory model to improve speech coders," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 411–414, 1993.

[58] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimising digital speech coders by exploiting masking properties of the human ear.," *Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[59] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[60] G. Theile, G. Stoll, and M. Link, "Low bit-rate coding of high quality audio signals," in *Proceeding of the 82nd Audio Engineering Society Convention*, pp. 158–181, August 1988.

[61] E. Ambikairajah, A. G. Davis, and W. T. K. Wong, "Auditory masking and mpeg-1 audio compression," *Electronics and Communication Engineering Journal*, pp. 165–175, August 1997.

[62] I. 11172-3, "Information technology- coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbits/sec," *ISO International Standard*, September 1993.

[63] B. K. Chua, *A Study of the MPEG-1 Audio Coding Standard*. MSc Thesis, The University Of Edinburgh, 1996.

[64] P. Noll, "Mpeg digital audio coding," *IEEE Signal Processing Magazine*, pp. 59–81, April/May 1997.

[65] K. Brandenberg and M. Bosi, "Overview of mpeg audio: Current and future standards for low-bit-rate audio coding," *Journal of the Acoustical Society of America*, vol. 45, pp. 59–81, January/February 1997.

[66] S. Nandkumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints on an auditory-based spectrum"," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 22–34, December 1995.

[67] J. H. L. Hansen and S. Nandkumar, "Robust estimation of speech in noisy background based on aspects of the auditory process," *Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3833–3849, 1995.

[68] S. Nandkumar and J. H. L. Hansen, "Speech enhancement based on a new set of auditory constrained parameters.," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 1–4, 1994.

[69] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using perceptual criteria," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 359–366, 1993.

[70] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1593–1602, 1994.

[71] A. A. Azirani, R. L. B. Jeannes, and G. Faucon, "Optimising speech enhancement by exploiting masking properties of the human ear," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 800–803, 1995.

[72] N. Virag, "Speech enhancement based on masking properties of the human auditory system."," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 796–799, 1995.

[73] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Transactions on Signal Processing*, vol. 39, pp. 1943–1954, September 1991.

[74] D. E. Tsoukalas, J. N. Mourjopolous, and G. Kokkinakis, "Speech enhancement based on audible noise suppression.," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 497–514, November 1997.

[75] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system."," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 126–137, March 1999.

[76] W. S. Ching and P. S. Toh, "Enhancement of speech signal corrupted by high acoustic noise.," *Proceedings of the IEEE Region 10 Conference on Computer Communication, Control and Power Engineering*, pp. 1114–1117, 1993.

[77] I. Pollack, "Monoaural and binaural threshold sensitivity for tones and for white noise," *Journal of the Acoustical Society of America*, vol. 20, pp. 52–57, January 1948.

[78] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *Journal of the Acoustical Society of America*, vol. 40, pp. 963–978, December 1992.

[79] X. Yang, K. Wang, and A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, pp. 824–839, March 1992.

[80] M. R. Schroeder, "Models of hearing," *IEEE Proceedings*, vol. 63, pp. 1332–1351, September 1975.

[81] E. D. Boer, "Classical and non-classical models of the cochlea," *Journal of the Acoustical Society of America*, vol. 101, pp. 2148–2150, April 1997.

[82] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, pp. 224–238, April 1997.

[83] R. F. Lyon, "A computational model of filtering, detection and compression in the cochlea," *IEEE Proceedings*, pp. 1282–1285, 1982.

[84] S. Voran, "Observations on auditory excitation and masking patterns," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 206–209, 1995.

[85] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, pp. 47–65, January 1940.

[86] A. Alwan, "A perceptual metric for masking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 712–715, 1993.

[87] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, pp. 155–182, 1979.

[88] W. W. Chang and C. T. Wang, "A masking threshold-adapted weighting filter for excitation search," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 124–132, March 1996.

[89] T. Usagawa, M. Iwata, and M. Ebata, "Speech parameter extraction in noisy environment using a masking model," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 81–84, 1994.

[90] M. Terry, S. Renals, R. Rowher, and J. Harrington, "A connectionist approach to speech recognition using peripheral auditory masking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 699–702, 1988.

[91] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 2524–2527, 1988.

[92] M. H. Goldstein Jr, "Auditory periphery as speech signal processor," *IEEE Engineering In Medicine and Biology*, pp. 186–196, April/May 1994.

[93] D. A. Sanders, *Auditory Perception of Speech*. Prentice Hall Press, 1977, ISBN 0-13-052787-4.

[94] S. V. Vaseghi, B. P. Milner, and J. J. Humphries, "Noisy speech recognition using cepstral-time features and spectral-time filters," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 65–68, IEEE, 1994.

[95] A. O. Abdel, M. A. Mokhtar, and M. A. Ez-El-Arab, "Speech enhancement in the communication between vehicles.," in *39th IEEE Vehicular Technology Conference*, vol. 2, pp. 897–901, IEEE, 1989.

[96] A. Drygajlo and M. El-Maliki, "Spectral subtraction and missing feature modeling for speaker verification.," in *Proceedings of EUSIPCO*, vol. 1, pp. 335–338, 1998.

[97] W. M. Kushner, V. Goncharoff, C. Wu, V. Nguyen, and J. N. Damoulakis, "The effects of subtractive type speech enhancement/noise reduction algorithms on parameter estimation for improved recognition and coding in high noise environments.," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 211–214, 1989.

[98] E. B. George, "Single-sensor speech enhancement using a soft-decision/noise attenuation algorithm.," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 816–819, 1995.

[99] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise.," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 208–211, April 1979.

[100] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *Journal of the Acoustical Society of America*, vol. 33, p. 248, 1961.

[101] H. Fletcher, "Auditory patterns.," *Reviews Of Modern Physics*, vol. 12, pp. 47–65, 1940.

[102] M. C. Killion, "Revised estimate of minimum audible pressure: Where is the "missing 6db"?.," *Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1501–1508, 1978.

[103] M. A. Tuffy and D. I. Laurenson, "Background noise reduction for mobile telephony.," in *IMA International Conference on Mathematics in Communications*, 1998.

[104] J. Taboada, S. Feijoo, R. Balsa, and C. Hernandez, "Explicit estimation of speech boundaries.," in *IEE Proceedings: Science, Measurement and Technology*, vol. 141, pp. 153–159, May 1994.

[105] P. Pollàk, P. Sovka, and J. Uhlíř, "Explicit estimation of speech boundaries.," in *Proceedings of EUROSPEECH'93*, pp. 1073–1076, September 1993.

[106] G. S. Kang and L. J. Fransen, "Quality improvement on lpc processed noisy speech by spectral subtraction."," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 37, pp. 939–942, June 1989.

[107] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppression," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[108] L. Hoy, B. Burns, D. Soldan, and R. Yarlagadda, "Noise suppression methods for speech applications," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1133–1136, 1983.

[109] G. Whipple, "Low residual noise speech enhancement utilising time frequency filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 5–8, 1994.

[110] Z. Goh, K.C.Tan, and B.T.G.Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 287–292, May 1998.

[111] S. Haykin, *Adaptive Filter Theory, 2nd Edition*. New Jersey: Prentice Hall Publishing, 1991, ISBN 0130132365.

[112] W. H. Haykin, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery, *Numerical Recipes in C, The Art of Scientific Computing, 2nd Edition*. New York: Cambridge University Press, 1996, ISBN 0-521431085.

[113] M. A. Tuffy and D. I. Laurenson, "Estimating clean speech masking thresholds for perceptual based speech enhancement.," in *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 1999.

[114] J. V. Tobias, *Foundations of Modern Auditory Theory, Volume II*. New York: Academic Press, 1972, ISBN 012691902X.

[115] P. C. Loizou, M. F. Dorman, and V. Powell, "The recognition of vowels produced by men, women, boys and girls by cochlear implant patients using a six channel cis processor.," *Journal of the Acoustical Society of America*, vol. 103, pp. 1141–1149, February 1998.

[116] J. C. Junqua, "The influence of psychoacoustic and psycholinguistic factors on listener judgements of intelligibility of normal and lombard speech.," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 361–364, 1991.