



## ABSTRACT

Title of Thesis:           AUDIO DSP: TIME AND FREQUENCY VARYING  
                                  GAIN COMPENSATION FOR NON-OPTIMAL  
                                  LISTENING LEVELS

                                  Thorvaldur Einarsson, Master of Science, 2003

Thesis directed by:       Professor Carol Espy-Wilson  
                                  Department of Electrical and Computer Engineering

Although the human hearing system is very complex, several models exist that explain parts of the hearing system. This thesis uses one of these models, the contours of equal loudness, to make music played at low listening levels sound more like it does at the intended listening level.

The perceived frequency balance of music varies with the listening level. This is especially noticeable at low listening levels, where frequencies below 500Hz seem attenuated. Moreover, hearing perception exhibits non-linear dynamic range compression, most evident at low frequencies.

A system is designed where filter banks and power measurements estimate the time-varying power of low frequency parts of the audio signal. The time-varying power of each narrow frequency band is compared to the contours of equal loudness, and

changes made to get the same frequency balance as at the intended listening level.

The thesis covers the design, implementation and performance of this system.

AUDIO DSP: TIME AND FREQUENCY VARYING GAIN COMPENSATION  
FOR NON-OPTIMAL LISTENING LEVELS

by

Thorvaldur Einarsson

Thesis submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Master of Science  
2003

Advisory Committee:

Professor Carol Espy-Wilson, Chair  
Professor Shihab Shamma  
Professor Min Wu



*"There is only one correct volume level for any particular piece of music"*

Peter Walker, Quad Electronics.

## TABLE OF CONTENTS

<b>Introduction .....</b>	<b>1</b>
<b>Part 1. Fundamentals.....</b>	<b>3</b>
1.1. Audio and Psychoacoustics: .....	3
1.1.1. Frequency Selectivity of the Ear - The Auditory Filters .....	3
1.1.2. Sensitivity of the Ear at Different Frequencies and Intensities – The Contours of Equal Loudness .....	5
1.1.3. Reference Listening Levels for Music .....	8
1.2. Selection of Filters.....	10
1.2.1. Downsampling (Anti-Aliasing) Filter .....	10
1.2.2. Interpolation (Anti-Imaging) Filter .....	11
1.2.3. Filter Banks .....	15
1.3. Previous Work.....	23
<b>Part 2. System Design .....</b>	<b>25</b>
2.1. Analysis .....	25
2.2. Processing.....	29
2.3. Synthesis.....	33
2.4. Efficient Implementation.....	34

Computational Complexity .....	35
Processing Delay .....	39
<b>Part 3. System Performance.....</b>	<b>42</b>
3.1. Performance of the Filter Banks.....	42
3.2. Objective Measurements .....	43
3.3. Subjective Measurements.....	45
<b>Conclusion.....</b>	<b>48</b>



## LIST OF TABLES

Table 1: Preferred listening levels for three groups, values show peak SPL values, data from [6].	9
Table 2: Filter specifications for anti-aliasing filter.	11
Table 3: Filter specifications for anti-imaging filter.	15
Table 4: Filter specifications for $H_0(z)$ .	19
Table 5: Sub-bands and respective frequency ranges for four-band tree structured filter bank. $f_s$ is the sampling frequency of the input signal $x[n]$ .	22
Table 6: Sub-bands and respective frequency ranges for three-octave filter bank. $f_s$ is the sampling frequency of the input signal $x[n]$ .	22
Table 7: Frequency ranges for octave filter bank.	26
Table 8: Shuffling order for 16-band filter-bank.	29
Table 9, Causes of processing delay and their respective delays.	40
Table 10: Comparison table from listening test.	46

## LIST OF FIGURES

Figure 1: Bandwidth of Critical Bands (CB), Equivalent Rectangular Bands (ERB) and one-third octave filters as a function of center frequency. ....	4
Figure 2: The contours of equal loudness, data from [2]. Each curve has a unit of Phon, e.g. all intensities on the 60Phon curve have the same perceived loudness as a 60dB SPL sine at 1kHz. The threshold of hearing lies at ca. 5Phon.....	6
Figure 3: Deviation from linearity in Phon curve spacing at 50Hz and 500Hz. ....	7
Figure 4: Magnitude spectrum for anti-aliasing filter .....	11
Figure 5: Mean squared error for a selection of interpolation filters, minimum is achieved for $\alpha=0.9$ and $L=10$ .....	12
Figure 6: Magnitude response of the optimal filter found from Figure 5 versus a filter based on the anti-aliasing filter as specified in Table 2. ....	13
Figure 7: Comparison of output waveforms from the two different anti-imaging filters.....	14
Figure 8: Two-channel QMF filter bank. ....	15
Figure 9: Magnitude response of analysis filters for filter bank. ....	20
Figure 10: Deviation from spectral flatness for the VSPM filter bank. ....	20
Figure 11: A four-band binary tree structured filter bank. ....	21
Figure 12: A three-octave tree structured filter bank. ....	22
Figure 13: Analysis part of system. ....	25
Figure 14: Comparison of sub-band width to Equivalent Rectangular Bands. ....	27

Figure 15: Example of compensation within a sub-band, numbers match steps 1 to 3  
in text.....30

Figure 16: Processing part of system for one frame.....31

Figure 17: Synthesis part of system. ....33

Figure 18: Four-level octave filter bank, showing the different sampling rates. ....36

Figure 19: Overview of the processing stage. ....37

Figure 20: Efficient structure for synthesis filters using type-2 polyphase  
decomposition. ....38

Figure 21: Top: Waveform before and after processing for a listening level of 80dB.  
Bottom: The 4096 point windowed DFT magnitude spectra of both signals. ....43

Figure 22: Top: Waveform before and after processing for a listening level of 100dB.  
Bottom: Corresponding DFT magnitude spectra. ....44

Figure 23: Histogram for listening test results, .....47

## **Introduction**

Human perception of sound is very complex and, despite considerable research activity in the area, is only understood to a limited degree. Hearing research has shown that our loudness perception is dependent on both frequency and intensity in a complex manner. For instance there is considerable dynamic range compression at low frequencies due to a higher threshold of hearing at these frequencies. This range compression is non-linear, as there is more compression at low intensities. Thus human perception of bass and low midrange is dependent on the sound intensity, both overall and for a given frequency at a given time. The result is that the perceived frequency balance of a recording or reproduction of sound will vary with the listening level. There will only be one “correct” listening level, the level of the original performance. This is most notable at low listening levels, where due to the non-linear dynamic range compression, bass and lower midrange sounds will seem attenuated.

Traditional methods have tried to overcome this problem by using filters to add gain at low frequencies when listening at low levels. These methods do not consider the dynamic range compression and its non-linearity and are often characterized by a “boomy”, unnatural sound.

The approach in this thesis is to design a system using filter banks and power measurements to estimate the time-varying power of low frequency parts of the audio signal. The power of a given frequency band at a given time is then compared to a model of loudness perception, and changes made to the signal accordingly.

Part 1 of this thesis will cover the fundamentals: psychoacoustics and hearing models, filters for downsampling and upsampling, basic structures of filter banks, as well as previous work. Part 2 covers system implementation, computational complexity, and possibilities for more efficient implementations. Part 3 looks at the performance of the system both in terms of objective measurements such as the mean square error, and more subjective measurements achieved by listening tests.

## **Part 1. Fundamentals**

This section will introduce basic concepts that are needed for this project as well as lay foundations for system structure and selection of parameters in part 2.

### **1.1. Audio and Psychoacoustics:**

In order to construct a system that makes changes to an audio signal in a manner that sounds natural to a listener, information on how the human hearing system works is needed. For the system implemented here, the three most important aspects of the human hearing system and music reproduction are:

1. Frequency selectivity of the ear, provides information on how to select filters that have similar characteristics.
2. Sensitivity of the ear at different frequencies and intensities provides a hearing model that can be used to determine the gain changes in each processing frame.
3. Reference listening levels for music have to be established to make changes to music at other listening levels according to the hearing model.

#### **1.1.1. Frequency Selectivity of the Ear - The Auditory Filters**

Early psychoacoustic experiments [1] have shown that the human auditory system acts as a spectrum analyzer. The basilar membrane in the inner ear acts as a series of bandpass filters, often referred to as the auditory filters [2]. Sounds that are picked up

by different parts of the basilar membrane have little effect on each other. But a sound that falls close in frequency to another sound can either become inaudible or render the other sound inaudible; this effect is called frequency masking and is widely used in audio compression. The frequency selectivity of the auditory filters has been estimated in many experiments and results vary with the methods used. Figure 1 shows a comparison of two auditory filter models, the Critical Bands (CB) as measured by Zwicker in 1961, and the more recent Equivalent Rectangular Bands (ERB) as measured by Glasberg and Moore in 1990. The figure also contains the bandwidth of a one-third octave filter, which is often used as an approximation of these models.

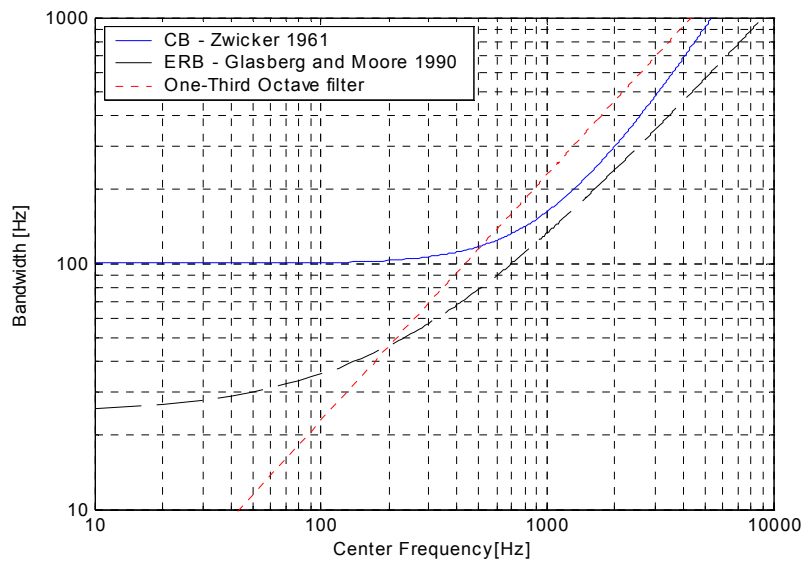


Figure 1: Bandwidth of Critical Bands (CB), Equivalent Rectangular Bands (ERB) and one-third octave filters as a function of center frequency.

There is some debate in the psychoacoustics community as to which model is better and both have their proponents. As the ERB model gives a narrower bandwidth, it is suitable as a conservative model and will be used in part 2 when the bandwidth of

sub-bands is considered. Moore [2] gives the following model of ERB bandwidth for a center frequency of  $f$ :

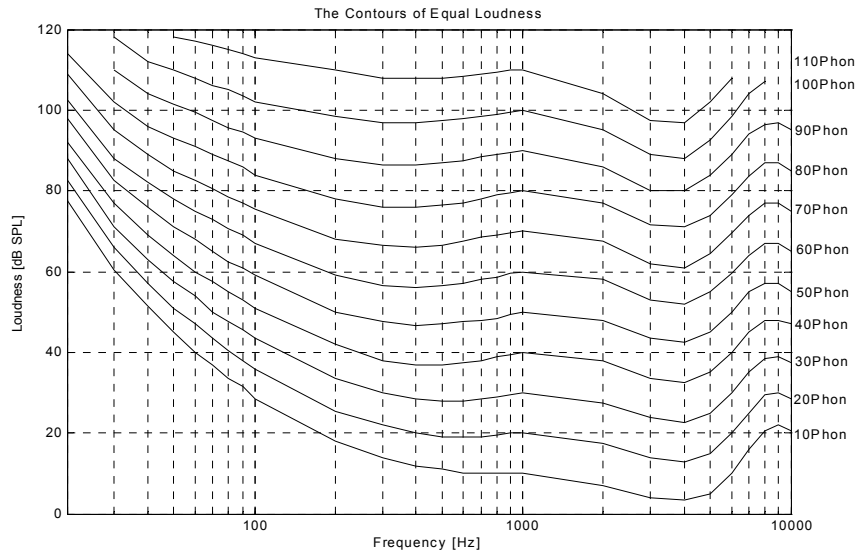
$$ERB = 24.7 \left( 1 + \frac{4.37f}{1000} \right) \quad [\text{Hz}] \quad (1)$$

Due to the constant factor in the equation, the bandwidth of ERB never goes below 24.7Hz as can be seen in Figure 1 where the bandwidth stays relatively constant for the low octaves and then increases more rapidly with frequency. Thus the one-third-octave filter whose bandwidth is a linear function of frequency is a poor approximation of the ERB at low frequencies.

### **1.1.2. Sensitivity of the Ear at Different Frequencies and Intensities – The Contours of Equal Loudness**

Research on ear sensitivity to different frequencies and intensities dates back to the research of Fletcher and Munson [3] in the 1930's. They conducted experiments where participating listeners were told to adjust the gain of a test tone to make it sound “as loud” as the reference, a 1kHz sine wave. By doing this over a variety of listeners, frequencies and intensities, a graph of equal loudness levels similar to that of Figure 2 was achieved.





*Figure 2: The contours of equal loudness, data from [2]. Each curve has a unit of Phon, e.g. all intensities on the 60Phon curve have the same perceived loudness as a 60dB SPL sine at 1kHz. The threshold of hearing lies at ca. 5Phon.*

Note that these experiments are done using sinusoids and thus can only give a broad picture of the more complex sounds that occur in music. Other psychoacoustic properties such as time and frequency masking [2] [4] will also play a role for complex sounds, but to which degree is impossible to say.

Figure 2 shows that in the frequency range of ca. 500-10,000Hz a change in sound intensity causes a similar change in perceived loudness level. However, at low frequencies the relation between intensity change and loudness level change is not as simple. This is partly due to the fact that substantially more power is needed to hear those frequencies, but also due to dynamic range compression and the non-linearity seen in the unequal Phon curve spacing. Figure 3 shows the deviation from linearity of Phon values at two frequencies, 50 and 500Hz. Data with 10Phon spacing is used and compared to the least squares linear estimation for each frequency.

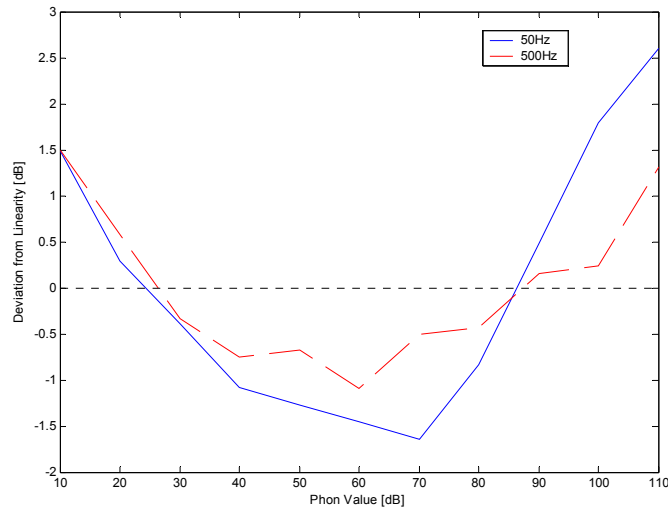


Figure 3: Deviation from linearity in Phon curve spacing at 50Hz and 500Hz.

Due to the definition of the Phon scale, at 1kHz, the Phon values are a linear function of the SPL values, but as Figure 3 shows, this is not true for other frequencies. Note also that the non-linearity increases at lower frequencies and that most of the dynamic range compression occurs at low intensities.

From the contours of equal loudness, it can thus be seen that changes in SPL have a bigger perceived effect at low frequencies than corresponding changes in the midrange and at high frequencies. The end result is that the perceived frequency balance of a recording or a reproduction of sound will vary with the listening level. There will only be one “correct” listening level, which is the level of the original performance. This will be most easily noticeable when listening at a low level, as low frequencies seem attenuated.

Another important observation is that some sounds that are heard at a higher listening level can fall below the threshold of hearing at low levels. This happens irrespective of frequency, and can be partly fixed by dynamic range compression that pushes

those sounds that should be heard, above the threshold of hearing. However, sounds at higher levels also need added gain, resulting in an altered frequency balance. This is an interesting research topic, but finding a compression curve for different frequencies that sounds natural is a considerable task. This research is thus limited to overcoming the unequal curve spacing of the Phon curves to restore the frequency balance of the audio at low listening levels. Part 2 will go in detail into how sounds at a low listening level can be made to sound more like they do at a higher level.

### **1.1.3. Reference Listening Levels for Music**

In order to make changes to music for low listening levels according to the hearing model above, a reference listening level for music must be established. In the case of acoustical recordings such as orchestral music, this level is simply the same as would be measured in the auditorium. For music that is more produced, such as pop music, the reference listening level can best be estimated as the level that the producer uses for mixing, most often quite loud.

Experiments have shown that the threshold of discomfort lies around 100dB SPL at most frequencies [5], so a normal dynamic range for music is between the threshold of hearing and the threshold of discomfort, perhaps with peaks reaching above 100dB SPL.

Slot [6] gives measurements from concert halls where intensities range from 40dB SPL for a very soft pianopianissimo (*ppp*) passage to 100dB for a strong fortifortissimo (*fff*) and further estimates that in the middle of a concert hall, peak sound intensity reaches about 100dB SPL. Slot cites another experiment where different groups of different people were asked to adjust the volume of music to the

level they prefer. It shows that musicians and sound engineers tend to listen to music at higher volume than the general public. Table 1 shows a part of the results of these experiments. Note however that these experiments date from the 1940's and both equipment and listening tastes have changed a lot since then.

*Table 1: Preferred listening levels for three groups, values show peak SPL values, data from [6].*

	Public		Musicians	Sound Engineers
	Men	Women		
Symphonic music	78	78	88	88
Light music	75	74	79	84

The difference in preferred listening levels between sound engineers and the public is notable and probably still exists today. The public thus does not listen to music at the reference level, which as mentioned above is the level used by the sound engineer and the artists for mixing and production. Note also that these levels are for dedicated music listening, whereas most of the time people have music playing in the background at a much lower level.

To further investigate the dynamic range of music, it is helpful to look at the Compact Disc Digital Audio (CDDA) format. CDDA uses 16bit PCM, which has a theoretical signal to quantization error ratio of 98dB [7]. This limits the dynamic range of the CD to 98dB.

Thus, in conclusion: A peak listening level of 100dB SPL will be used as a reference level from here on, or equivalently, a sine wave using the full amplitude of 16bit PCM gives 100dB SPL at the reference listening level.

## **1.2. Selection of Filters**

Digital filters are used for three purposes in this project:

1. A downsampling (anti-aliasing) filter separates the frequency range to be processed from the higher frequencies that are not to be altered.
2. An interpolation (anti-imaging) filter is used after expanding the processed samples to remove images caused by the expander.
3. Filter banks that split the signal into narrow frequency bands to prepare for processing, and then after processing gather the sub-bands for output.

This section summarizes important properties of filters for all three uses and then goes on to the selection of appropriate filters.

### **1.2.1. Downsampling (Anti-Aliasing) Filter**

The role of the downsampling filter is to reduce the frequency range before decimation to avoid aliasing artifacts. As will be shown in part 2, the incoming audio is a 16 bit PCM signal, sampled at 44.1kHz, and needs to be downsampled by a factor of 32 to isolate the frequency band of 0-689Hz for filter bank analysis. For 32-fold downsampling for high quality audio use, the anti-aliasing filter must have a very narrow transition band (<200Hz), the passband ripple must be small (<0.5dB), and stopband attenuation must be high (>60dB). As linear phase is needed to avoid phase distortion, an FIR filter meeting these demands must have a high order and thus requires a lot of processing. The filter shown in Table 2 and Figure 4 was designed using Parks-McClellan algorithm [8], and chosen as it gives a good trade-off between filter length and performance.

Table 2: Filter specifications for anti-aliasing filter.

Description	Symbol	Value
Filter Order	$N$	500
Passband Edge	$F_{pass}$	480Hz
Stopband Edge	$F_{stop}$	$f_s/64 \cong 689\text{Hz}$
Passband Ripple	$A_{pass}$	0.3dB
Stopband Attenuation	$A_{stop}$	60dB

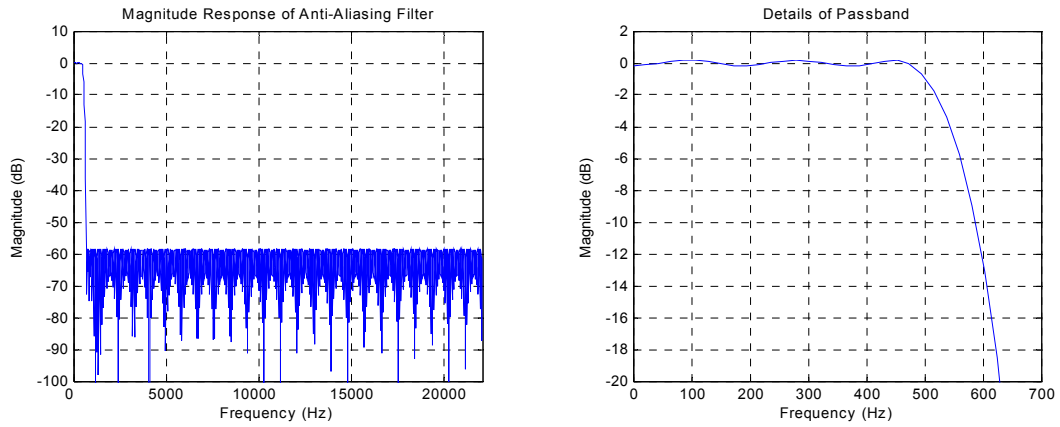


Figure 4: Magnitude spectrum for anti-aliasing filter

As the high frequency part of the input signal needs to be retained by subtracting the low-pass signal from the input before downsampling, it is not possible to use polyphase decomposition, Interpolated FIR (IFIR) [9] or other such computationally efficient schemes that involve decimation as intermediary steps for the decimation filter. Efficient implementations will be dealt with in a special sub-chapter.

### 1.2.2. Interpolation (Anti-Imaging) Filter

The purpose of the anti-imaging filter is to remove frequency images created by the expander. It is a lowpass filter that meets the same requirements as the anti-aliasing

filter specified above, except it needs added gain of 32 (30.1dB) in the passband to make up for gain loss due to the removal of spectral images. Thus the same filter design should do quite well. However, experimental results showed a different filter selection for the anti-imaging filter could attain an output waveform much closer to the input waveform. Matlab has a built in interpolation function [10]

$$Y = INTERP(X, R, L, \alpha) \quad (2)$$

that expands the input signal  $X$  by a factor  $R$  and then applies an anti-imaging filter specified by parameters  $L$  and  $\alpha$ , where:

$\alpha$  is the normalized cutoff frequency of the input signal,  $0 < \alpha \leq 1$ .

$L$  is specified as half the number of original sample values used to perform the interpolation. Ideally  $L$  should be less than or equal to 10.

Figure 5 shows the Mean Square Error (MSE) for a number of values of  $0 < \alpha \leq 1$  and  $L \leq 10$  using a relatively short music segment.

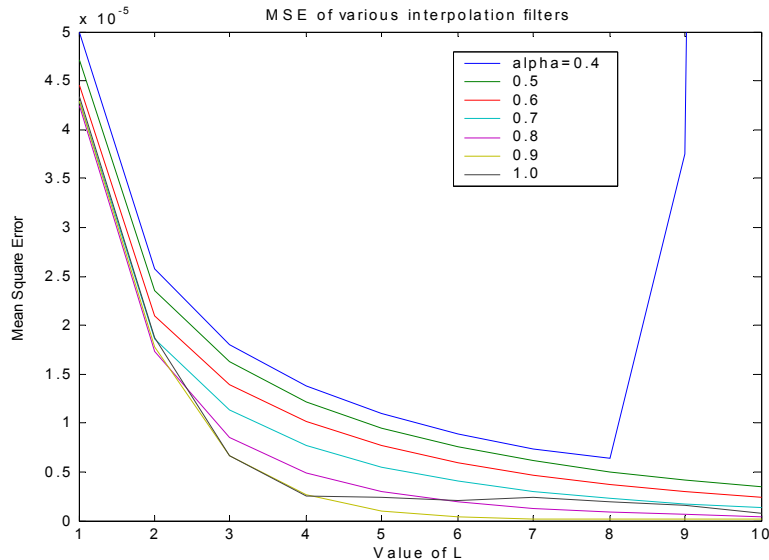


Figure 5: Mean square error for a selection of interpolation filters, minimum is achieved for  $\alpha=0.9$  and  $L=10$ .

The lowest MSE is found to be  $1.3197 \cdot 10^{-7}$  for values  $\alpha=0.9$  and  $L=10$  resulting in a filter order of  $N=640$ . By comparison a filter identical to the anti-aliasing filter gives a considerably higher MSE of  $2.2383 \cdot 10^{-5}$  for  $N=500$ . Higher values of  $L$  were tried, but they gave only slightly better results and were not used as the filter gets a higher order, and also because such high values of  $L$  are not recommended by the Matlab documentation [10]. Figure 6 shows a comparison of the optimal filter from Matlab and the anti-aliasing filter with added gain.

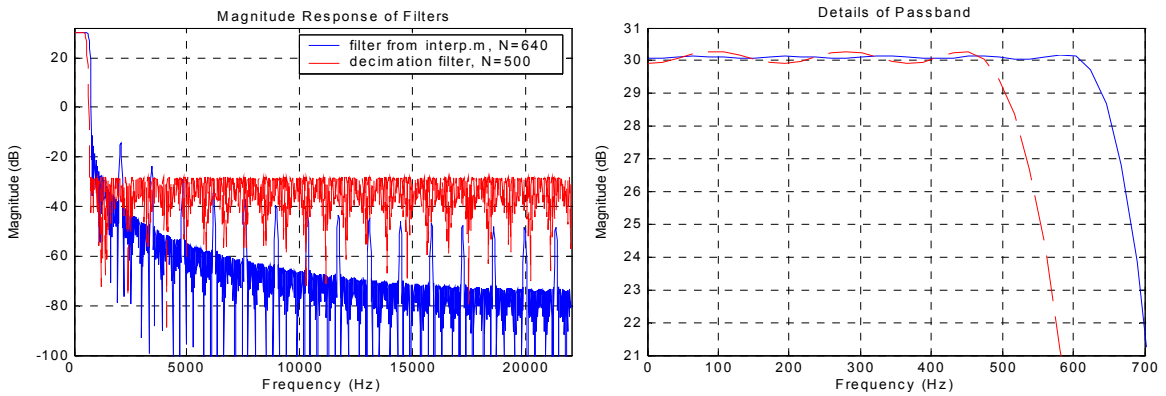


Figure 6: Magnitude response of the optimal filter found from Figure 5 versus a filter based on the anti-aliasing filter as specified in Table 2.

Comparing the two magnitude responses, the optimal filter found for the interpolation function in Matlab has generally poorer stopband attenuation. This is especially true close to the passband where it achieves only 8dB attenuation at the stopband edge,  $F_{stop} = 689\text{Hz}$ . However the attenuation increases with frequency and at high frequencies the attenuation becomes considerably greater than for the anti-aliasing filter. The anti-aliasing filter has better overall stopband attenuation (>60dB over the whole stopband), but more ripples in the passband.



The difference in MSE and the reason the Matlab interpolation filter does better is most likely due to smaller passband ripples as well as the superior attenuation at higher frequencies.

Figure 7 shows a sample waveform with anti-imaging done by both filters. It can be seen that the filter from Matlab gives an output closer to the original signal.

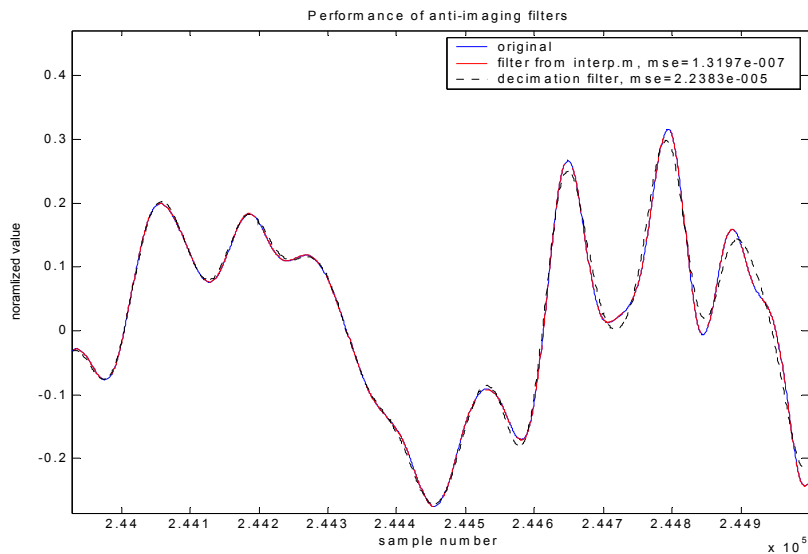


Figure 7: Comparison of output waveforms from the two different anti-imaging filters.

Thus the filter from Matlab was selected as the anti-imaging filter for this project. Its specifications are given in Table 3.

Table 3: Filter specifications for anti-imaging filter.

Description	Symbol	Value
Filter Order	$N$	640
Passband Edge	$F_{pass}$	610Hz
Stopband Edge	$F_{stop}$	770Hz
Passband Ripple	$A_{pass}$	0.15dB
Stopband Attenuation	$A_{stop}$	>40dB

### 1.2.3. Filter Banks

Two-Channel Filter Banks:

Following is a short introduction on filter bank characteristics. For a complete discussion see [9]. Figure 8 shows a basic filter bank structure, the two-channel Quadrature Mirror Filter (QMF) bank.

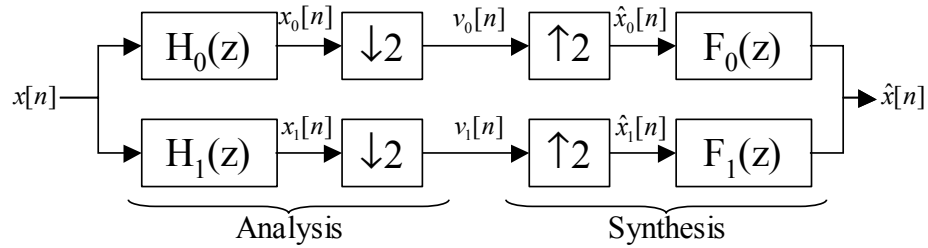


Figure 8: Two-channel QMF filter bank.

The input  $x[n]$  is filtered by  $H_0(z)$  and  $H_1(z)$ , where  $H_0(z)$  is typically lowpass and  $H_1(z)$  highpass such that its magnitude response  $|H_1(\omega)|$  is the mirror image of  $|H_0(\omega)|$  with respect to the center frequency  $\pi/2$ , hence the name Quadrature Mirror Filter bank [11].

$$H_1(z) = H_0(-z) \Rightarrow |H_1(\omega)| = |H_0(\omega - \pi)| \quad (3)$$

The outputs of filters  $H_0(z)$  and  $H_1(z)$  are then decimated two-fold to give the subbands,  $v_0[n]$  and  $v_1[n]$ .

The synthesis part starts by two-fold expansion, and follows with filtering by  $F_0(z)$  and  $F_1(z)$ . The filter outputs are then summed together to give the output signal  $\hat{x}[n]$ .

Z-domain analysis of this system gives the following input/output relationship:

$$\hat{X}(z) = T(z)X(z) + A(z)X(-z), \quad (4)$$

where

$$T(z) = \frac{1}{2}[H_0(z)F_0(z) + H_1(z)F_1(z)] \quad (5)$$

is called the distortion transfer function [9] and

$$A(z) = \frac{1}{2}[H_0(-z)F_0(z) + H_1(-z)F_1(z)] \quad (6)$$

can be called the aliasing transfer function.

A filter bank is said to have Perfect Reconstruction (PR) if

$$\hat{x}[n] = \alpha x[n - \beta] \Leftrightarrow \hat{X}(z) = \alpha z^{-\beta} X(z) \Leftrightarrow T(z) = \alpha z^{-\beta}, \quad (7)$$

for some constants  $\alpha$  and  $\beta$ ,  $\alpha \neq 0$ . This guarantees that the output is the same as the input with the exception of constant gain and delay terms, thus with no amplitude, phase or aliasing distortion.

It is very important to eliminate aliasing distortion, and that can be achieved by making sure  $A(z) = 0$ . This is most often done by choosing the synthesis filters as

$$F_0(z) = H_1(-z) \text{ and } F_1(z) = -H_0(-z), \quad (8)$$

causing aliasing terms that are generated by the expander to cancel each other out when summed together for the output.

To avoid phase distortion the filter bank needs to have linear phase, achieved by requiring all analysis and synthesis filters to be linear phase FIR.

Finally, amplitude distortion for an alias free filter bank is avoided by satisfying a spectral flatness criterion:

$$|H_0(\omega)|^2 + |H_1(\omega)|^2 = 1 \text{ for } \omega = [0, 2\pi]. \quad (9)$$

Note that these choices limit the possible number of filter bank designs. However, all filters can be derived from  $H_0(z)$ , which simplifies the design process.

It turns out that with the above selection of filters, perfect reconstruction restricts the choice of filters so much that the resulting filters have very poor performance.

Another way of achieving perfect reconstruction is by not enforcing  $H_1(z) = H_0(-z)$ , and instead choosing  $H_0(z)$  to be power symmetric, that is

$$H_0(z)\tilde{H}_0(z) + H_0(-z)\tilde{H}_0(-z) = 1, \quad (10)$$

where

$$\tilde{H}(z) \equiv H^*(z^{-1}). \quad (11)$$

Then by choosing

$$H_1(z) = -z^{-N}\tilde{H}_0(-z), \quad (12)$$

perfect reconstruction can be achieved with the appropriate selection of the synthesis filters [9]. The resulting filter bank can have good attenuation, but is rather tricky to design and  $H_0(z)$  does not have linear phase.

To overcome difficulties and limitations in the design of perfect reconstruction filter banks, very often so-called near-perfect reconstruction is used. For near perfect reconstruction, aliasing and phase distortion are eliminated, but some amplitude

distortion is allowed. The aim is then to minimize the amplitude distortion by ensuring

$$|H_0(\omega)|^2 + |H_1(\omega)|^2 \cong 1. \quad (13)$$

Several methods have been devised that optimize the filters to give good stopband attenuation while still meeting the spectral flatness criterion. The most commonly used technique is that of Johnston [12], but more recent methods have shown better results [13].

#### Appropriate Filter Bank Selection:

For the filter banks used in this project, the following properties are most important:

1. Steep attenuation outside passband: Processing causes considerable changes to the sub-band signals by altering gain according to the psychoacoustic model. This can cause two adjoining sub-bands to have quite different gain. Even if the filter bank is aliasing free, the difference in gain will introduce some aliasing, as the aliasing terms do not cancel out as they would if the gain were the same in both bands. The steeper the transition band and more attenuated the stopband, the less aliasing is allowed, reducing the effect of gain changes on distortion.
2. Near perfect reconstruction: It is important to add little or no distortion to the signal due to the filter bank selection. Perfect reconstruction seems a good choice, but near perfect reconstruction banks can have linear phase analysis filters and very steep transition bands, thus satisfying property 1.

To meet the two properties above, an order 64 linear phase equiripple FIR filter from [13] was chosen. The filter is designed using the Vector Space Projection Method (VSPM). It has near perfect reconstruction and performs better for a given order than the classic Johnston filters [9]. The specifications and magnitude response of  $H_0(z)$  can be seen in Table 4 and Figure 9 respectively.  $H_1(z)$ ,  $F_0(z)$  and  $F_1(z)$  are all derived from  $H_0(z)$  using

$$H_1(z) = H_0(-z), \quad (14)$$

$$F_0(z) = 2H_0(z) \text{ and} \quad (15)$$

$$F_1(z) = -2H_0(-z). \quad (16)$$

Maximum deviation of  $|H_0(\omega)|^2 + |H_1(\omega)|^2$  from unity is only 0.005dB and is shown in Figure 10.

Table 4: Filter specifications for  $H_0(z)$ .

Description	Symbol	Value
Filter Order	N	63
Passband Edge	$F_{pass}$	$0.425\pi$
Stopband Edge	$F_{stop}$	$0.595\pi$
Passband Ripple	$A_{pass}$	0.003dB
Stopband Attenuation	$A_{stop}$	71dB

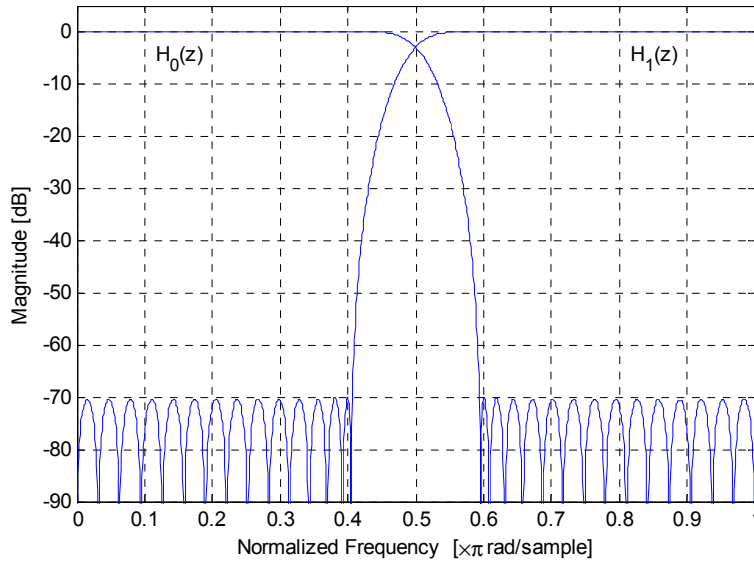


Figure 9: Magnitude response of analysis filters for filter bank.

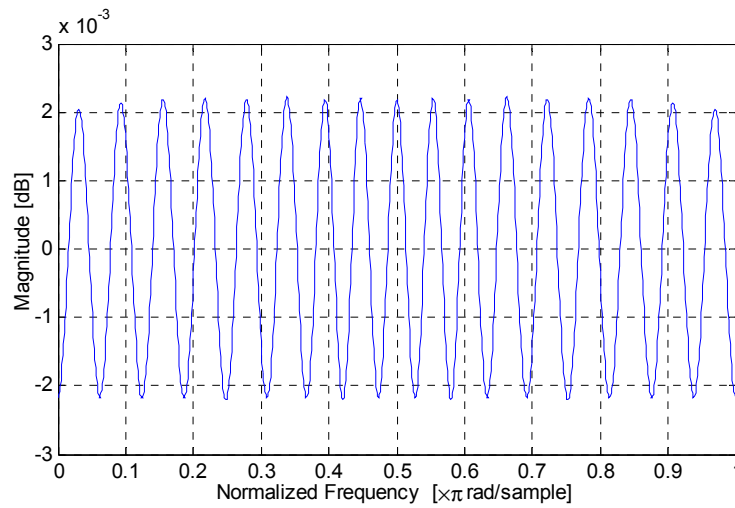


Figure 10: Deviation from spectral flatness for the VSPM filter bank.

### Multi-Channel Filter Banks:

The two-channel filter bank chosen above only divides the signal into two frequency bands. For the purposes of this project, very narrow sub-bands over a limited range of the audible spectra are needed, and thus multi-channel filter banks that meet these requirements must be developed.

A powerful yet convenient method of designing multi-channel filter banks is by repeated use of the two-channel bank. These are referred to as tree structured filter banks [9], [14]. Figure 11 shows a four-band, or two-level, binary tree structured filter bank.

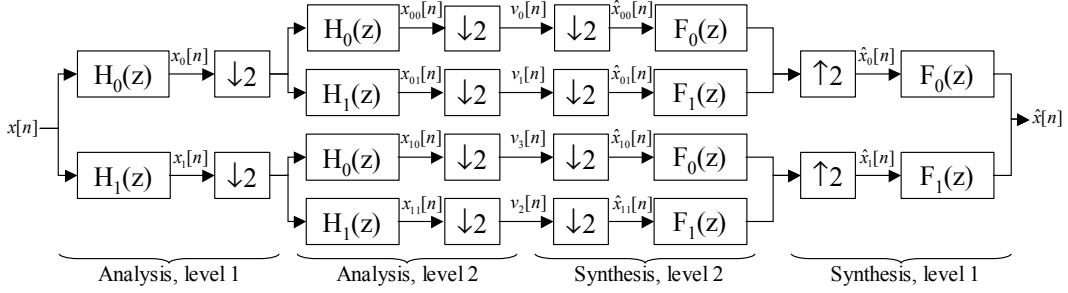


Figure 11: A four-band binary tree structured filter bank.

Generally different levels of the tree structured bank can have different filter sets, but usually the filters used are the same at all levels. Two-channel filter bank properties such as perfect reconstruction and alias free conditions also hold for a tree based filter bank if they hold for each level [9]. Note that, since  $H_1(z)$  is a highpass filter,  $x_1[n]$  has its energy concentrated in the high frequencies  $[f_s/4, f_s/2)$ , where  $f_s$  is the sampling frequency of  $x[n]$ . This causes the downsampling to reverse the frequency of the signal. Due to this frequency reversal, further filtering with  $H_1(z)$  gives a lower frequency sub-band than filtering with  $H_0(z)$ . Thus  $x_{11}[n]$  becomes sub-band  $v_2[n]$  after decimation and  $x_{10}[n]$  becomes  $v_3[n]$ . The frequency ranges of the resulting sub-bands are shown in

Table 5.



Table 5: Sub-bands and respective frequency ranges for four-band tree structured filter bank.  $f_s$  is the sampling frequency of the input signal  $x[n]$ .

Sub-Band Name	Frequency Range	Frequencies
$v_0[n]$	$[0, f_s/8)$	Normal
$v_1[n]$	$[f_s/8, f_s/4)$	Reversed
$v_2[n]$	$[f_s/4, 3f_s/8)$	Normal
$v_3[n]$	$[3f_s/8, f_s/2)$	Reversed

Tree structured filter banks can allow for flexibility in sub-band width by applying more filtering to some sub-bands than others, often called multiresolution filter banks [9]. An example of that can be seen in Figure 12, which shows a three-octave filter bank.

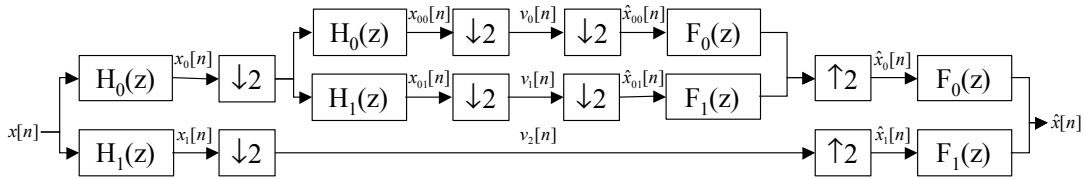


Figure 12: A three-octave tree structured filter bank.

Table 6 shows the frequency range of the resulting sub-bands.

Table 6: Sub-bands and respective frequency ranges for three-octave filter bank.  $f_s$  is the sampling frequency of the input signal  $x[n]$ .

Sub-band Name	Frequency Range	Frequencies
$v_0[n]$	$[0, f_s/8)$	Normal
$v_1[n]$	$[f_s/8, f_s/4)$	Reversed
$v_2[n]$	$[f_s/4, f_s/2)$	Reversed

Note that not all sub-bands are octaves. The two higher sub-bands,  $v_1[n]$  and  $v_2[n]$  represent each a doubling of frequency, whereas the lowest sub-band,  $v_0[n]$  in fact covers infinitely many octaves. The two lowest sub-bands in an octave filter bank will always have the same bandwidth and sampling rate. In order for an octave filter bank to meet conditions such as perfect reconstruction or alias-free, again those conditions must hold for each level of the filter bank. This means that if there is any distortion in the two-channel sub-bank then the distortion function must be added to the higher octave sub-bands. For the filter bank in Figure 12; if the two-channel filter bank that yields sub-bands  $v_0[n]$  and  $v_1[n]$  has distortion transfer function  $T(z)$ , then sub-band  $v_2[n]$  must be filtered by  $T(z)$  for aliasing cancellation of the whole system. Both of the tree structures mentioned above can be expanded to more sub-bands by adding levels of two-channel filter banks.

### **1.3. Previous Work**

As the contours of equal loudness have been known since the 1930's, many attempts have been made to overcome this perceptual effect. The "Loudness Control" button found on most stereo systems is generally an analog bass boost circuit, often with varying bass boost depending on the listening level [15]. Often the high frequencies are boosted too [16], probably with the intention of making the music sound more "exciting" as the effect of the threshold of hearing can make music sound somewhat "dull". As these systems are only variable by listening level, they cannot handle the non-linear dynamic range compression of the equal loudness curves at low

frequencies. The resulting sound is thus known to sound unnatural and “boomy” and is generally avoided by discerning listeners.

Modern stereo and home theater systems are using digital signal processing more and more, but the algorithms used in these systems are proprietary and information is hard to obtain. The best way to find out about such things is a patent search, which in this case turned up empty. To the author’s best knowledge, a system such as implemented here has not been implemented before.

## Part 2. System Design

This section covers system design specifications. The three stages of the system, analysis, processing and synthesis will be discussed first, followed by a discussion on computational complexity and possibilities for a real-time implementation of the system.

### 2.1. Analysis

The role of the analysis section of the system is to take the input audio signal and prepare it for processing by splitting it into 80 sub-bands. As only frequencies up to 689Hz will be processed, the analysis section must also retain the high frequency part of the signal.

Input data is regular CD data; 16bit, 44.1kHz PCM and stereo channels are treated separately. Specifications for individual filters were found in part 1. Figure 13 shows the analysis part of the system as implemented.

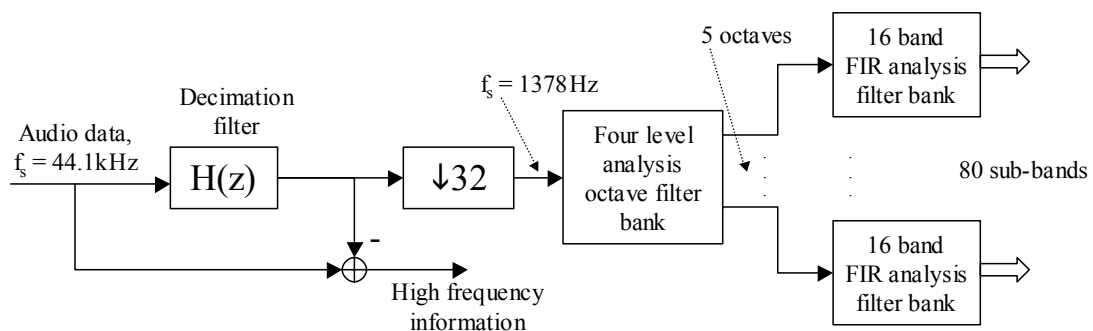


Figure 13: Analysis part of system.

The decimation filter  $H(z)$  removes high frequency components from the signal before decimation. Before the decimation the output of the filter is subtracted from

the delayed input signal to retain the high frequency part of the original signal. The high frequency information will be added to the processed low frequency part during synthesis. After decimation the sampling frequency is 1378Hz and that signal is fed into a four level octave analysis filter bank. By using a regular two-band QMF bank structure and repeating the filtering for the low-pass part only, the outputs will be five separate octaves in the range of 0 to 689Hz. The filter bank uses the order 63 VSPM FIR filter (selected in part 1 as the prototype for the two-band QMF filter bank) and is implemented using the discrete wavelet transform. Note however, as the filters are not wavelet bases, the bank can at best be called a wavelet like filter bank.

Each of the five octaves is next fed into a sixteen band, equally spaced filter bank. This filter bank also uses the order 63 VSPM FIR. The 16 sub-bands are generated using a tree structure [9] [14], where both the low and high pass outputs become inputs to the next stage. The resulting total number of sub-bands in the range of 0-689Hz is 80. Table 7 shows the resulting frequency ranges as well as bandwidth of the sub-bands.

*Table 7: Frequency ranges for octave filter bank.*

Octave Number	Frequency Range of Octave	Sampling Frequency	Sub-band Width of 16-band Filter Bank
1	0-43Hz	86Hz	2.7Hz
2	43-86Hz	86Hz	2.7Hz
3	86-172Hz	172Hz	5.4Hz
4	172-344Hz	344Hz	10.77Hz
5	344-689Hz	689Hz	21.53Hz

As Table 7 shows, the sub-bands are spaced quite close in the lower octaves. The lowest octave, 0-43Hz is not really an octave, since its bandwidth is more than a doubling of frequency. However as the human hearing system can only hear frequencies down to about 20Hz, the frequency range of the sub-band is approximately equal to the lowest octave of human hearing. Notice that the two lowest octaves have the same width of sub-bands, as they are both outputs from the same level of the filter bank. This turns out to fit the Equivalent Rectangular Bands (ERB) model discussed in part 1 quite well, as can be seen in Figure 14, where the bandwidths of the sub-bands are compared to the ERB model.

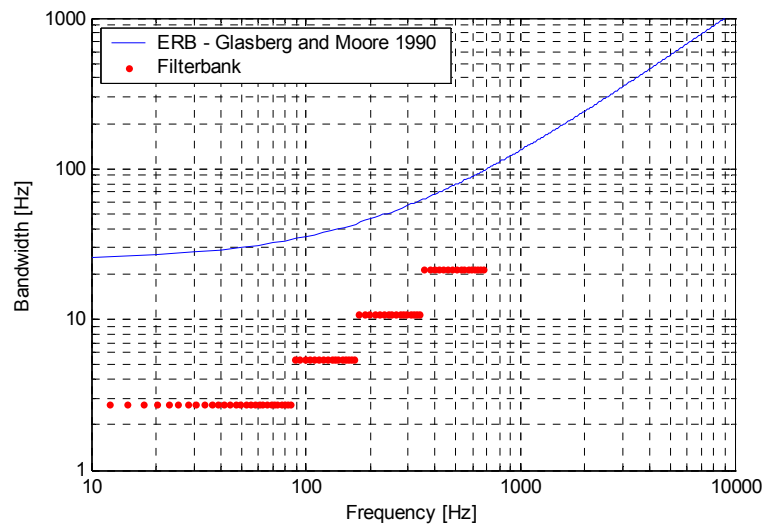


Figure 14: Comparison of sub-band width to Equivalent Rectangular Bands.

All sub-bands have bandwidths well below the frequency selectivity of the ear according to the ERB model. In fact, 8 sub-bands per octave would also have been below the ERB curve and 4 sub-bands per octave would touch the curve at some points. The reasons for choosing 16 sub-bands are the following:

1. 16 sub-bands allow finer frequency resolution for the processing part of the system.
2. As the signals in adjoining sub-bands are more related than in the 8 sub-band case, it is likely that there is less difference in processing between adjoining sub-bands, resulting in reduced aliasing distortion.
3. The extra sub-bands are computationally cheap to implement, as the data rate is already very low.
4. The filters used in the filter bank have very steep attenuation outside the passband and their frequency overlap is minimal. The ear on the other hand can be considered to have a lot of filters with significant overlap, each with bandwidth corresponding to the ERB. Sounds occurring within an ERB centered on a masking tone will be masked. However, the masker must be at the center of the band and for non-overlapping filters that is not possible if the filters have bandwidth equivalent to one ERB.

As mentioned in part 1, tree based filter banks affect the order of sub-bands in relation to the frequencies they contain due to frequency reversal. Thus before comparison to a psychoacoustic model that is dependent on the frequency range of the sub-band, the sub-bands must be shuffled. Table 8 shows the shuffling order, note that the order is different for the 16-band filter stemming from lowest octave, as it is the output of a series of lowpass filters, whereas the other 16-band filters stem from a high pass that causes a frequency reversal.

Table 8: Shuffling order for 16-band filter-bank.

Sub-band # Low to High Freq.	Shuffle Lowest Octave	Shuffle Other Octaves	Sub-band # Low to High Freq.	Shuffle Lowest Octave	Shuffle Other Octaves
1	1	9	9	13	5
2	2	10	10	14	6
3	4	12	11	16	8
4	3	11	12	15	7
5	7	15	13	11	3
6	8	16	14	12	4
7	6	14	15	10	2
8	5	13	16	9	1

## 2.2. Processing

The role of the processing stage is to alter the sub-band signals given by the analysis stage to make music played back at a low level sound more like it sounds at the reference level. This is done by comparing the power of the sub-band signals to the equal loudness curves introduced in part 1 and deriving the optimal power for each sub-band at a given time and listening level.

### Demonstrative Example

In order to understand better what the processing part does, it is beneficent to start with an example that gives a general idea about the processing. Figure 15 shows the equal loudness contours for an imaginary narrow sub-band centered at 70Hz.



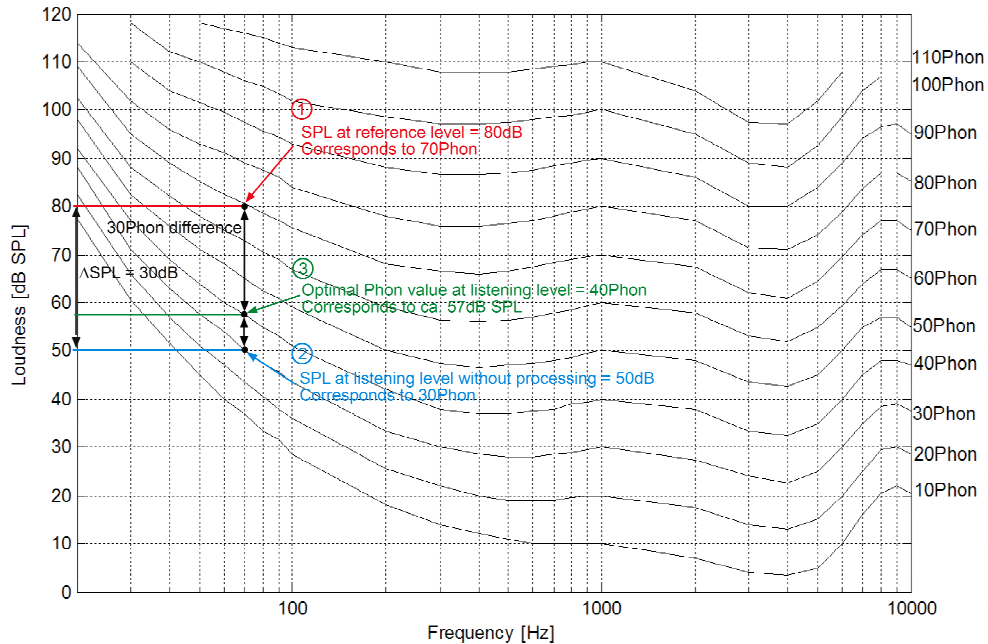


Figure 15: Example of compensation within a sub-band, numbers match steps 1 to 3 in text.

The processing can be divided into three steps:

- Step 1. The sub-band is first divided into short frames for power measurement relative to the reference level of 100dB. This gives the Sound Pressure Level (SPL) in the frame when listening at the reference level. Using the approximation that the Phon value is constant within a sub-band, the SPL value is then converted to its corresponding Phon value. In this example the power measurement at the reference level gives 80dB SPL, which lies on the 70Phon curve.
- Step 2. Let  $\Delta\text{SPL}$  be the difference in dB between the reference listening SPL and the actual listening SPL. In this example the listening level is 70dB SPL, so  $\Delta\text{SPL} = 30\text{dB}$ . Without processing, the signal in the sub-band would be heard at  $80\text{dB} - 30\text{dB} = 50\text{dB}$  SPL ( $\Delta\text{SPL}$  below the measured value at the

reference listening level). This level corresponds to 30Phon, and is thus 40Phons lower than when listening at the reference level. For the correct frequency balance at this listening level relative to higher frequencies, this sub-band should sound at 40Phon, or  $\Delta\text{SPL} = 30\text{Phons}$  below the reference level.

Step 3. The SPL value corresponding to the optimal Phon value, 40Phon, is now found, resulting in a value of ca. 57dB SPL. Thus, now we have the SPL value for which the signal in the sub-band sounds relatively equally loud as it would at the reference listening level. The difference between the optimal SPL value and the SPL value without processing is  $57\text{dB} - 50\text{dB} = 7\text{dB}$ . Thus the sub-band needs added gain of 7dB to achieve the correct frequency balance.

### Detailed Description

Figure 16 shows a block diagram of the systems processing part:

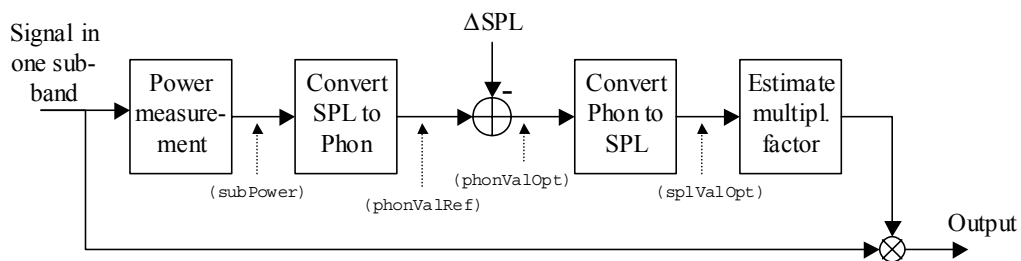


Figure 16: Processing part of system for one frame.

The inputs are the 80 sub-bands from the analysis part. Each sub-band is split into short frames and the power of each frame at the reference listening level measured in decibels using:

$$subPower = round\left(P_{ref} + 10 \log_{10}\left(\frac{1}{k} \sum_{n=1}^k x^2[n]\right)\right) \quad [\text{dB}] \quad (17)$$

Here  $P_{ref}$  is the reference listening level in dB, selected in part 1 as 100dB,  $x[1]$  is the first sample in the frame and  $k$  is the frame length. This power measurement thus constitutes the SPL that would be measured within that frequency band if listening at the reference level. The framelength  $k$  should be as small as possible so that changes in power are detected, yet big enough so that the measurement gives an accurate reading for all the frequencies within the sub-band. The value of  $k$  was chosen as  $k=8$  and with no overlap between frames, giving a good tradeoff between accuracy and response time.

The value from the power measurement is then converted to  $phonValRef$ , the corresponding Phon value, using table lookup. The table was created from the equal loudness contours in [2], but as the graph had only values every 10Phon, the data was interpolated in both axes to a resolution of 1dB SPL and to the center of each sub-band.

To get the optimal level in Phons we must subtract  $\Delta\text{SPL}$ , the difference between the reference and the listening level, from the reference Phon value

$$phonValOpt = phonValRef - \Delta\text{SPL}. \quad (18)$$

This is logical because at higher frequencies, a change of 10dB SPL is also a change of 10Phon and so we have the Phon values as they should optimally be for the given listening level.

The Optimal Phon value,  $phonValOpt$  is now converted to  $splValOpt$ , the optimal SPL level using a table derived from the same data as the SPL to Phon table

described above. This table maps values from Phons to SPL at the center of the sub-bands with a resolution of 1Phon.

Finally the multiplication factor for each frame is calculated by using

$$SplValList = subPower - \Delta SPL, \quad (19)$$

$$multFactor = 10^{(splValOpt - splValList)/20}, \quad (20)$$

where  $SplValList$  is the SPL of the frame when listening without any processing. The multiplication factor is then used to adjust the gain of the frame by being multiplied with the contents of the frame.

### 2.3. Synthesis

The role of the synthesis part is to take the processed sub-band signals and convert them back to the same format as the system's input signal. Figure 17 shows the structure of the system's synthesis part:

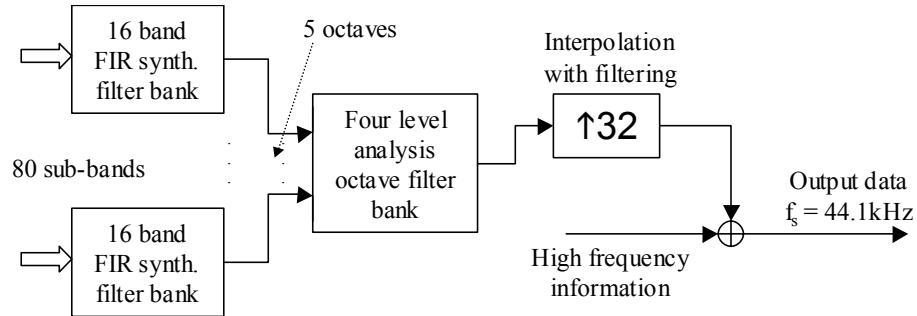


Figure 17: Synthesis part of system.

The synthesis is basically the reverse of the analysis, using corresponding synthesis filters. The 80 sub-bands are first joined to 5 octaves and then those octaves are joined to form the output signal for the frequency range of 0-689Hz. The octaves go through different numbers of filters due to the non-symmetrical tree construction,

resulting in different group delay for different octaves depending on the number of filters they go through. It is thus important for phase synchronization to add delay in some octaves so that all octaves have the same amount of delay due to filtering.

The output of the filter banks has a sampling frequency of 1378Hz and thus needs to be expanded by 32 to be sampled at 44.1kHz. The order 640 anti-aliasing filter selected in part 1 successfully removes images due to the expansion. This signal is finally summed up with a delayed version of the high frequency information to give a phase-synchronized output. This is possible as all the filters used in the system are linear phase FIR and thus have constant group delay. Output data is of the same format as the input, 16bit PCM at 44.1kHz sampling rate.

## **2.4. Efficient Implementation**

The purpose of this research is to check the viability of using psychoacoustic models to make changes to audio so that the listening experience is improved at low listening levels. The main incentives are thus to get as good frequency resolution and as little distortion as possible. As the system is implemented in Matlab, the input audio file must be treated all at once using matrix operations, not allowing for real-time operation. Therefore, other performance issues such as efficient filter implementation, processing delay and memory use are unimportant. However, as these issues are all important for a real-time implementation, it is beneficial to look at how the system performs, and what needs to be changed in order to make the system realizable and efficient in real-time.

### 2.4.1. Computational Complexity

Here, the computational complexity of an efficient implementation of the system will be estimated by looking at possible computation savings due to the structure of filters, as well as the complexity of the processing stage. As both stereo channels are treated identically, the calculations here are done for one channel only. The measure used for computational complexity will be Multiplications Per Unit-time (MPU). This measurement gives the number of multiplications needed for one input sample, and is based on the input rate of the system (44.1kHz) that must be taken into account when measuring the complexity of filter working at a lower rate.

#### Analysis Stage

The anti-aliasing filter that precedes the decimator (see Figure 13) is a real coefficient, linear phase lowpass FIR filter of order 500. Thus it is symmetric, such that  $h[n] = h[N-n]$ . This symmetry allows for using only  $\lceil (N+1)/2 \rceil = 251$  multipliers instead of  $N+1 = 500$  multipliers needed for a direct form implementation of one sample [9]. Further reduction of complexity is not possible as mentioned earlier, due to the fact that the output of the filter must be subtracted from the high frequency before decimation by 32.

Both the octave and 16-band filter banks use filters based on  $H_0(z)$ , so  $H_1(z) = H_0(-z)$ .

As  $H_0(z)$  is a symmetric odd order filter,  $H_1(z)$  is anti-symmetric. For the octave analysis filter bank, the two lowest octaves require the most computations, as the samples must go through four two-channel filter banks. Each two-channel analysis filter can be implemented using type-1 polyphase decomposition where

$H_0(z) = E_0(z^2) + z^{-1}E_1(z^2)$  and  $H_1(z) = E_0(z^2) - z^{-1}E_1(z^2)$ , where  $E_0(z)$  and  $E_1(z)$  are called the polyphase components [9]. This allows for very efficient implementation by filtering after the decimation and using similarities between  $E_0(z)$  and  $E_1(z)$  due to the symmetry of  $H_0(z)$ . The total number of multiplications needed for each stage is thus just  $(N+1)/4$  for each input sample. However, the sampling rates at the inputs vary depending on what stage of the tree-based filter the filtering is being done at. Figure 18 shows the polyphase implementation as well as the corresponding sampling rates at the different stages of the filter bank.

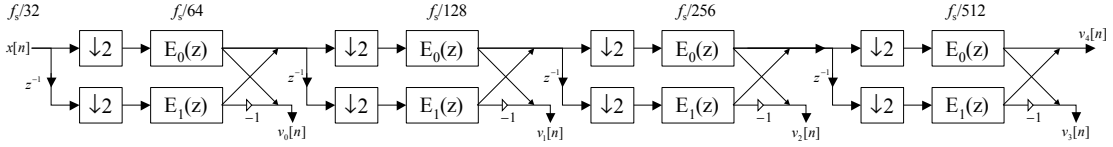


Figure 18: Four-level octave filter bank, showing the different sampling rates.

The number of multiplications needed for octave  $i$  is

$$MPU_{octave}(i) = \frac{1}{32} \frac{N+1}{4} \sum_{k=0}^i \frac{1}{k^2}, \quad i=0, \dots, 4 \quad (21)$$

Each of the outputs  $v_0[n], \dots, v_4[n]$  then goes through the 16-band analysis filter bank that can also be implemented using polyphase decomposition. In this case all resulting sub-bands go through four levels of filters. Each level of the tree-based 16-band filter bank needs the same number of multiplications as each filter needs half the number of multiplications of a filter in the preceding level, but there are twice as many filters relative to the preceding level. If the input to the 16-band filter bank has sampling frequency  $f_{s(i)}$ , where  $i=0, \dots, 4$  is the number of the octave it belongs to, then the number of multiplications needed for each 16-band filter bank is

$$MPU_{16\text{-band}}(i) = \frac{4f_{s(i)}}{f_s} \frac{(N+1)}{4} = \frac{(N+1)}{32(i+1)}. \quad (22)$$

The total MPUs are thus

$$MPU_{analysis}(i) = \sum_{i=0}^3 MPU_{16\text{-band}}(i) + MPU_{octave}(i) \quad (23)$$

The results for  $N=63$  and  $i=0, \dots, 4$  are 7.2MPUs. This is negligible considering that the decimation filter requires 251MPUs and is achieved because the sampling rate for these filters is so much lower.

### Processing Stage

As the filter banks are maximally decimated, each sample that is input to the analysis filter banks results in one output from the filter banks. Thus it suffices to look at the number of calculations needed for each output from the filter banks. Figure 19 shows again a schematic diagram of the processing stage.

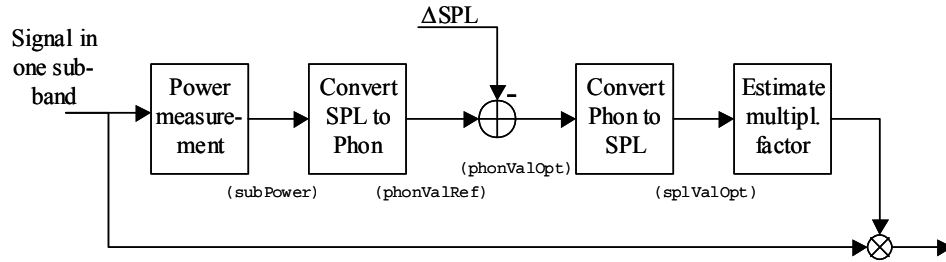


Figure 19: Overview of the processing stage.

The amount of calculations per input sample the processing stage needs is negligible. First of all, the sampling rate at the input of the filter banks is  $f_s/32$  and each output of the filter bank is grouped with 8 other samples in a frame. The operations following the power measurement are only done once for each frame. Conversions of SPL to Phon and Phon to SPL are done using table lookup and are thus cheap to implement.



One multiplication is needed to adjust the gain of each sample, and the power measurement can be done by squaring each value, summing up and dividing by the framelength before converting to decibels. The total number of operations depends on how efficiently logarithms and non-integer powers can be treated, but due to the low rate at the input of this stage, the total number of MPUs (relative to the input sampling rate of 44.1kHz) needed is close to one.

### Synthesis Stage

The filter banks in the synthesis stage can be implemented using a similar structure as the polyphase filters in the analysis stage, this structure is shown in Figure 20. The filters can thus be implemented using the same amount of MPUs as the analysis stage (7.2MPUs).

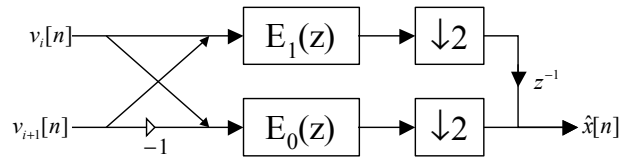


Figure 20: Efficient structure for synthesis filters using type-2 polyphase decomposition.

The anti-imaging filter that follows the expander can be implemented much more efficiently than the anti-aliasing filter in the analysis stage. The reason for this is that there is no need to do the filtering at a full rate, as was needed for the anti-aliasing filter. Thus, the filtering can be done using polyphase decomposition and utilizing the symmetry within the polyphase components [9]. This implementation requires

$$MPU_{anti-imaging} = \frac{N+1}{64} \cong 10 \quad (24)$$

for each output, given 32-fold expansion and a filter length of  $N=640$ .

## Total Computational Complexity

The total computational complexity of the system is bound by the complexity of its sub-parts. From the numbers above, the total number of multiplications per input sample can be as low as ca. 280MPUs (1.25 million operations per second) for one channel.

Note that a vast majority of these multiplications is due to the decimation filter in the analysis part and this suggests a different approach to the decimation filter should be taken for a more efficient implementation, even though it might cause some distortion to the high frequency part of the input signal.

### **2.4.2. Processing Delay**

The high frequency resolution and sharp cutoffs that are achieved in the filter banks used in this project come at the cost of processing delay. The delay of an  $N$ -th order linear-phase FIR filter is the same as the group delay, which is constant at  $\lceil N/2 \rceil$ .

The delay due to the processing stage is the number of samples used in the power measurement. The maximum delay for the analysis and synthesis banks, as well as the processing stage occur for the lowest frequency octaves that have a sampling rate of 5.4Hz. Table 9 shows the various processing delays in the system.

Table 9, Causes of processing delay and their respective delays.

Cause of Delay	Delay in samples	Sampling Frequency	Delay in Time
Decimation Filter	250	44100Hz	5.7ms
Analysis Filter Banks	32	5.4Hz	5.9s
Processing	8	5.4Hz	1.5
Synthesis Filter Banks	32	5.4Hz	5.9
Interpolation Filter	320	44100Hz	7.3ms

The delays due to the decimation and interpolation filters are several orders of magnitude smaller than the delays in the filter banks. The filter bank delay of 5.9s is due to the very low sampling rate of the lowest sub-bands and the resulting total processing delay of over 13s is not acceptable in a real-time implementation. However, to reduce this, frequency resolution must be sacrificed. By not implementing the 16-band filter bank for the lowest frequency bands, the filter bank delay goes down to 0.37s and by doing the power measurement with overlap, the delay due to frames can also be reduced. Thus it seems possible, without great sacrifices to quality, to achieve a total processing delay of less than one second, but there always exists a tradeoff between frequency resolution and processing delay.

## 2.5. Possible Extensions

Many other uses are possible for this kind of system to improve the listening experience under non-optimal circumstances. An example is listening in the presence of noise, such as in cars, where road noise masks soft sounds in a frequency dependent manner. Thus adaptive frequency dependent dynamic range compression

based around filter banks similar to those developed here would give much better results than simple waveform range compression that is often used in such circumstances.

## Part 3. System Performance

### 3.1. Performance of the Filter Banks

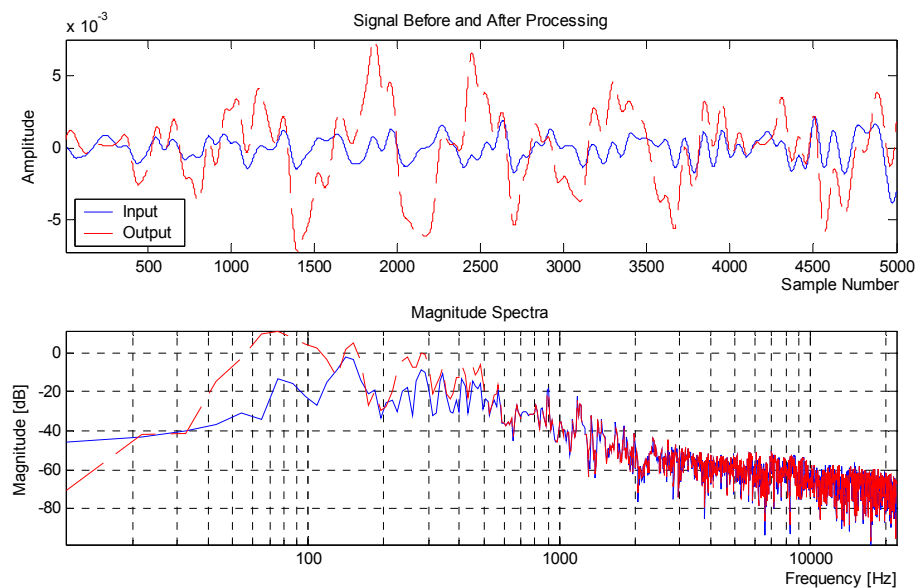
By bypassing the processing part of the system, the performance of the filter bank can be measured. The Mean Square Error (MSE) is calculated and converted to decibels relative to the power of the input signal.

$$MSE_{dB} = 10 \log_{10} \left( \frac{\sum_n (\hat{x}[n] - x[n])^2}{\sum_n (x[n])^2} \right) \quad (25)$$

Here,  $x[n]$  is the input to the filter bank (the downsampled version of the input audio signal), and  $\hat{x}[n]$  is the output of the filter bank. The resulting average  $MSE_{dB}$  over various different sound files is inaudible,  $-68\text{dB}$ . This measurement can also be done for the input audio signal and the output audio signal, thus including the decimation and interpolation filters. The performance for this measurement is somewhat worse, but still quite acceptable, giving an average of  $-49\text{dB}$ . Thus a lot of the distortion is due to the decimation and interpolation filters. As these filters already take up a substantial part of the computations needed for the system, increasing their length to make them sharper is not advisable. Another option might be to directly filter the input signal using an octave filter bank that would cover all octaves up to  $f_s/2$ . However, as the filter bank used does not have perfect reconstruction, this would also add some error to the output.

### 3.2. Objective Measurements

Figure 21 shows a comparison between an input audio signal and the corresponding processed output for a listening level of 80dB. It is evident from the magnitude spectra, that for frequencies above 689Hz the two signals are the same. This was to be expected as the high frequency information is kept without changes and added to the processed low frequency output with correct delay. At low frequencies however, the difference is considerable. This can be seen in the waveform that retains its shape at high frequencies, but the low frequencies are clearly reinforced. Also noticeable is that there is no phase distortion at any frequency.



*Figure 21: Top: Waveform before and after processing for a listening level of 80dB. Bottom: The 4096 point windowed DFT magnitude spectra of both signals.*

As the system makes deliberate changes to the input signal at low frequencies, measures such as the mean square error do not give any information about the performance of the system in normal operation. However, there is one more

measurement where the MSE gives some information, that is by setting the listening level equal to the reference level. As the power measurements cannot be totally accurate, and the Phon to SPL and SPL to Phon tables have limited precision, the result is a worse mean square error than for no processing. The average  $MSE_{dB}$  over several different sound files measured between the output and input signals was  $-30\text{dB}$ . To see this, Figure 22 contains both the waveforms and magnitude spectra for the case where both listening, and reference levels are  $100\text{dB}$ . Clearly there is some difference in the input and output signals, but listening tests have shown that there is no audible difference in the two.

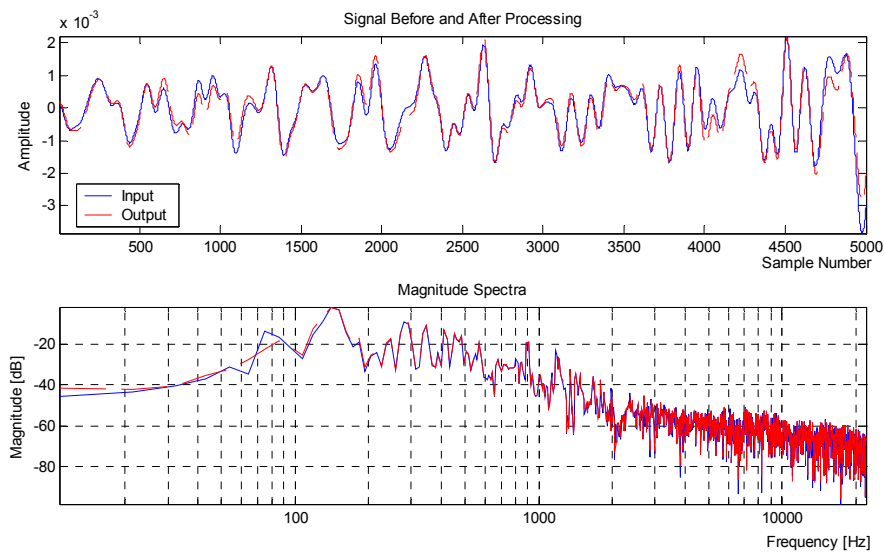


Figure 22: Top: Waveform before and after processing for a listening level of  $100\text{dB}$ . Bottom: Corresponding DFT magnitude spectra.

For other listening levels,  $MSE_{dB}$  has no significance due to the changes made. We must therefore rely mainly on subjective measurements.

### **3.3. Subjective Measurements**

This system has been tested subjectively, by comparing the processed and original versions using either high quality headphones or stereo systems. Informal listening tests have given very good results and all listeners have favored the processed version over the non-processed version when listening at a low listening level. To get more concrete results, a formal listening test was performed with a number of listeners and blind comparison.

Usually such listening tests are done using pairs. Then, two different versions of the same test track are played to the listener that has to judge the quality. The quality is most often rated on a scale of 1 to 5, the result of which is called an opinion score. For this system, the method of comparing two versions of the same track is not good. This was tried using two attenuated versions of the test track, one processed and one unprocessed, and the listener was told to say which sounded better. It turned out that the listener's preference for music, especially concerning the level of bass was too big a factor, so a second, more neutral test was devised.

The test was set up in the following manner:

Listening examples of length 10-15 seconds were taken from five different pieces of music, chosen to reflect many genres of music and contain notable low-frequency content. These examples were processed using a reference level of 100dB and a listening level of 80dB. This level is a typical level music playing in the background or listening at night. An audio CD was made where each musical piece is represented in three versions. The first version is not processed and at full volume (100dB). The second and third versions contain either unprocessed at lower volume (80dB), or the



piece processed in system for a lower volume (80dB), selected in a random manner. Two second gaps were inserted between the tracks. For statistical accuracy and measurement of listener’s consistency, repeats of the same pieces in different orders would be interesting. But a listeners in the first test complained about fatigue due to listening to the same music again and again in random order, it was decided to do without repeats. This also made the listening test shorter, which helped getting volunteers.

A portable CD player and Grado SR-125 headphones were used for the comparison. Before starting the test, the playback level was calibrated using an SPL meter and a 1kHz sine wave. Listeners were mainly graduate students and all reported to have normal hearing. The listeners were not trained before starting the test. They were only told to listen to difference in the bass region. Each listener was then told to rate which one of the latter two versions sounds more like (truer to, closer to) the first version on a scale of one to five, as seen in Table 10.

*Table 10: Comparison table from listening test.*

Second is much closer	Second somewhat closer	About as close	Third somewhat closer	Third is much closer
1	2	3	4	5
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

The results of the listening test can be seen in Figure 23, where results for all five musical pieces are combined.



*Figure 23: Histogram for listening test results,*

In percentages: 60% say the processed version sounds much or somewhat closer to the original, 26% say both versions sound as close, and 14% say the unprocessed sounds much or somewhat closer to the original. Thus a significant majority prefers the processed version, giving the conclusion that the system meets expectations by improving the listening experience at low listening levels.

## **Conclusion**

A system has been designed and constructed that aims at improving the listening experience when listening to music at a low listening level. This is achieved by using filter banks and comparison to a hearing model for narrow low frequency bands. Both informal qualitative listening tests, and blind comparison testing show that the system achieves these goals.

The system was designed with very good frequency resolution, which makes the processing delay unacceptably long for a real-time implementation. However by sacrificing some of the frequency resolution, this delay can be significantly reduced. A good tradeoff between quality and delay would need to be found for a real-time implementation. The computational complexity using efficient filter structures is about 280 multiplications per unit time, and the system can thus be implemented in real-time using modern DSP processors.

## References

- [1] Fletcher, H. 1940. *Auditory Patterns*, Rev. of Modern Physics, 12, pp. 47-65
- [2] Moore, B. 1997, *An Introduction to the Psychology of Hearing*, Academic Press, San Diego, CA.
- [3] Fletcher, H.; Munson, W.A. 1933. *Loudness, its Definition, Measurement and Calculation*, J. Acoust. Soc. Am. 9, pp. 82-108.
- [4] Zwicker, E. & Fastl, H. 1999, *Psychoacoustics, Facts and Models*, 2<sup>nd</sup> ed., Springer-Verlag, Heidelberg, Germany.
- [5] Poulsen, T. 2000. *Ear, Hearing and Speech*, Department of Acoustic Technology, Technical University of Denmark.
- [6] Slot, G. 1971. *Audio Quality*, Drake Publishers, New York, NY.
- [7] Pohlman, K.C. 2000. *Principles of Digital Audio*, 4<sup>th</sup> ed., McGraw-Hill.
- [8] Porat, B. 1997. *A Course in Digital Signal Processing*, Wiley.
- [9] Vaidyanathan, P.P. 1993. *Multirate Systems and Filter Banks*, Prentice Hall.
- [10] Matlab R13 help files. 2002, The Mathworks.
- [11] Strang, G. & Nguyen, T. 1996. *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA.

- [12] Johnston, J.D. 1980. *A Filter Family Designed for use in Quadrature Mirror Filter Banks*, IEEE ICASSP, April 1980, pp. 291-294.
- [13] Haddad, K.C.; Stark, H.; Galatsanos, P. 1998. *Design of Two-Channel Equiripple FIR Linear-Phase Quadrature Mirror Filters Using the Vector Space Projection Method*, IEEE Signal Processing Letters, July 1998, pp. 167-170.
- [14] Woods, J.W.; O'Neil, S.D. 1986. *Subband Coding of Images*, IEEE Trans on ASSP, No 5, Oct. 1986, pp. 1278-1288.
- [15] White, G.D. 1991. *The Audio Dictionary*, University of Washington Press, Seattle, WA.
- [16] Coulter, D. 2000. *Digital Audio Processing*, R&D Books, Lawrence KS.

