

SOUND RESYNTHESIS FROM A CORRELOGRAM

by

Daniel Naar

May, 1993

Technical Report No. 3
NSF Grant No. IRI-9214233

Department of Electrical Engineering
San Jose State University
San Jose, California 95192

Sound Resynthesis from a Correlogram

A Thesis
Presented to
The Faculty of the Department of Electrical Engineering
San Jose State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

By
Daniel Naar
December, 1993

Approved:

Professor Richard Duda
Thesis Advisor

Date

Graduate Committee:

Professor Rangaiya Rao

Date

Professor Udo Strasilla

Date

Professor Richard Duda

Date

Professor Peter Reischl

Date

Professor Belle Wei

Date

Abstract

A unique method of analysis creates a representation of sound, termed the correlogram, that may allow sounds produced simultaneously to be separated. If a separation technique is developed it will prove useful to be able to resynthesize the original sound. The resynthesis poses a difficult problem, because the analysis procedure includes non-linearities and removes phase information from the signal. In this thesis we integrate and modify a set of known techniques thereby developing an algorithm to resynthesize speech from its correlogram. The algorithm is implemented and shown to successfully resynthesize speech.

Acknowledgments

I am greatly indebted to many people who have provided technical direction, advice, and above all moral support. To Malcolm Slaney I send my heartfelt thanks for his support and unrelenting demand for excellence; his stiff but constructive criticism has helped me to produce a work of which I am truly proud. I want to thank Dick Lyon and Apple Computer Inc. for providing such an exciting and challenging environment in which to work. Both Bill Stafford and Shihab Shamma supplied me with excellent feedback and ideas. My advisor, Dick Duda, provided me ample warning of the pitfalls that I would encounter; unfortunately I did not realize the true wisdom of his advice till much later in this work. I do not have the eloquence to even begin to thank my family, so I will merely state the obvious, "Thank you for everything." Almost last, but definitely not least, I thank my loving partner and super sweetie, Annie, for putting up with me, proof-reading my very first and most terrible draft, and of course for just plain helping out. To everyone else that contributed, thanks! This work was partially supported by the National Science Foundation under grant number IRI-9214233.

TABLE OF CONTENTS

1. Introduction	1
2. Forward Processing—Cochleagram and Correlogram	4
2.1 Cochlear Model	6
2.1.1 Cochlear Filter	7
2.1.2 Automatic Gain Control (AGC)	8
2.2 Correlogram Calculation	9
3. Sound Waveform Estimation from Cochleagrams	12
3.1 Cochlear Filter Inversion	12
3.2 Inverting Half-Wave Rectification	18
3.3 Inverting Automatic Gain Control	27
4. Correlogram Inversion	30
4.1 Signal Estimation from its Short Time Fourier Transform Magnitude	30
4.2 Improving the Initial Guess	34
4.3 Including Knowledge of Signal Characteristics	37
5. Resynthesizing Speech	43
6. Conclusion	45
Bibliography	47

1 Introduction

A great deal of effort has been invested to understand the processing performed by the human auditory system. The auditory system is capable of performing extraordinary feats, such as adapting to inputs that span a 100 dB range, recognizing speech in very noisy environments, as well as separating and recognizing different sound sources that occur simultaneously. Beyond the purely scientific interest of understanding these capabilities, we believe that improved modeling of the auditory system will lead to better and highly robust front ends for speech recognition systems.

The auditory system converts sound from dynamic variations in pressure to neural activity and ultimately to complex representations in the brain. This metamorphosis begins in the cochlea where mechanical vibrations are transformed into motion of the basilar membrane. Due to the physical properties of the membrane and its associated receptor cells, the sound is distributed as a set of band-pass signals whose center frequencies are a function of position along the membrane. The receptors, called hair cells, are attached to the membrane and convert its movement into neural impulses which are transmitted, via the auditory nerve, to higher levels in the auditory system. In his duplex theory of pitch perception Licklider (1951) proposed that the outputs of the cochlea are autocorrelated, resulting in a representation that combines both frequency domain and time domain processing. Combining a mathematical model of the cochlea (Lyon, 1982) and a calculation similar to that proposed by Licklider, Lyon and Slaney have created a novel representation of sound termed the correlogram (Slaney and Lyon, 1993).

The correlogram, a three dimensional representation of sound, is displayed as a movie. Each frame is composed of the short-time autocorrelation functions of the cochlear model outputs. A set of frames from a correlogram of a segment of speech is shown in Figure 1. The vertical axis represents place along the cochlea (channel or frequency), and the horizontal axis represents the delay in the autocorrelation function; both time and frequency-domain information are preserved in this representation. Because the autocorrelation function is symmetric, only positive delays are shown. The intensity of the

correlation is proportional to the darkness in the image. The dark horizontal bands indicate those areas of the spectrum where the energy is most concentrated, much like a spectral estimate. In the voiced sounds, these bands indicate the location of the formants of the vowel. The vertical bands indicate the temporal periodicity in the signal, effectively emphasizing the pitch. For the unvoiced sound, the fourth frame of Figure 1, there is no pitch and distinct dark bands cannot be seen. Thus, the correlogram maintains a duplex representation of both time (periodicity) and frequency.

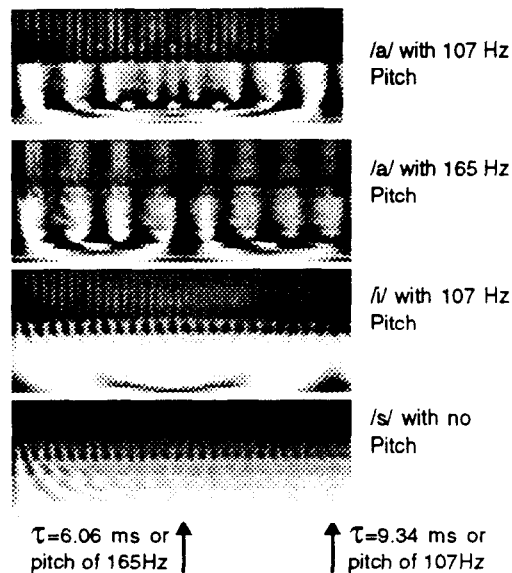


Figure 1. Four frames of a correlogram of speech. The first three frames are voiced sounds. When the pitch is raised the vertical structures become closer together (first and second frame). As the formant frequencies change the horizontal bands move (first and third frame). Finally, if the sound is unvoiced (last frame) then there is no vertical structure. Reprinted with permission (Slaney & Lyon, 1993)

The correlogram can be a very useful tool to visualize sound. A set of examples has been developed and made available on videotape (Slaney & Lyon, 1991) which allows the viewer to appreciate this representation of sound. It is not possible to display these movies here. However, a description of the correlogram of a complex sound composed of three synthesized vowels played simultaneously, /a/ (haa), /i/ (eek), and /u/ (boo) (Duda, Lyon and Slaney, 1990), may aid the reader to perceive the possible benefits of such a representation. The sound is initially created using a constant pitch for each vowel, which results in an irritating buzz and correlograms that are equally ambiguous. However, when the pitch of one of the vowels is frequency modulated by five percent, the modulated vowel emerges from the cacophony and can be easily perceived by the listener. A similar result

is experienced when viewing the correlogram. The varying pitch causes portions of the correlogram to modulate along the horizontal axis and an object appears to emerge which is highly correlated with the auditory sensation of the listener. Examples such as these have lead to the hypothesis that the correlogram may be a useful tool for isolating sounds.

A similar representation, named the coincidence function, was used by Weintraub (1985) to separate the voices of two speakers. The separation algorithm tracked the two dominant pitches and assigned each to a speaker. Furthermore, it determined the absence or presence of speech, and in the latter case decided whether the speech was voiced or unvoiced. The resulting pitch track for each speaker was then used to determine an amplitude ratio function for each channel of the cochlear model, which in turn allowed the separated speech to be resynthesized directly from the output of the cochlear filter.

A strikingly different approach to perform separation is to group objects that appear to move together in the correlogram domain. This is similar to associating many geese crossing the horizon as one flock or object. Associated objects are grouped together under the premise that they represent one sound source. The question then arises, however, as to how well such correlogram fragments separate sounds without losing critical information. The autocorrelation function is known to lose all phase information, and it is not obvious that a recognizable, let alone good quality sound can be resynthesized from a fragment of the correlogram.

This thesis explores the problem of resynthesizing a sound from its correlogram. Resynthesizing the original sound requires the inversion of both the cochlear model and the correlogram. The cochlear model inversion results in almost perfect results and no degradation in the sound quality. Inversion from the correlogram itself provides some quite surprising results. Not only is the sound quality barely degraded, but a time-based comparison can be made between the reconstructed signal and the original. These results show that sounds reproduced from the correlogram inversion are excellent reproductions.

Prior to examining the inversion, the forward processing necessary to create the cochleagram and correlogram is examined in the second chapter. The inversion problem

is divided into two smaller components. The first includes the issues present when inverting the cochleagram. In the third chapter the issues surrounding the estimation of a sound stream from its cochleagram are examined in detail. In the fourth chapter, the more difficult problem of estimating the cochleagram from the correlogram is addressed. Finally, the entire procedure is brought together in the fifth chapter to resynthesize sounds from their correlograms.

2 Forward Processing—Cochleagram and Correlogram

To compute the correlogram, sound is converted into neural firing rates using the cochlear model. Then a short-time autocorrelation of each channel is calculated to create the correlogram. A block diagram illustrating these processing steps is presented in Figure 2.

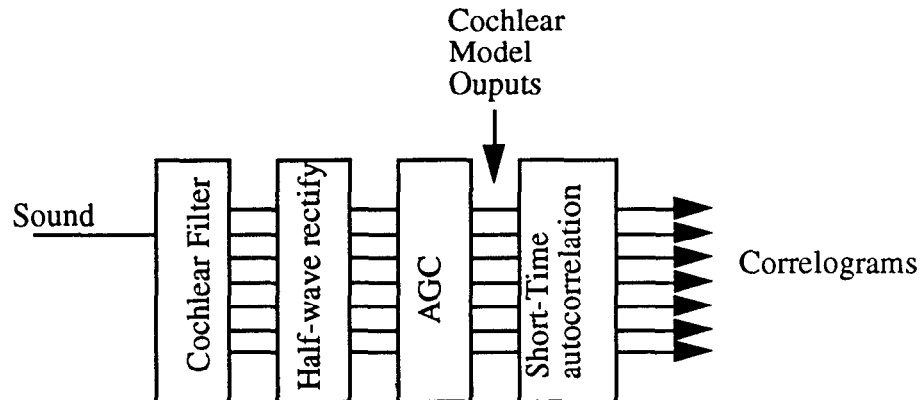


Figure 2. Correlogram computation block diagram.

A more graphical description is shown in Figure 3. The expansion of the dimensionality clearly can be seen as the representation of sound is converted from a one dimensional signal to the three dimensional correlogram representation.

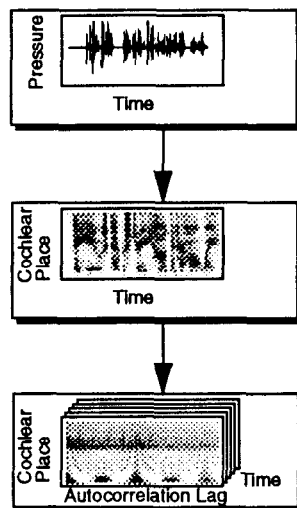


Figure 3. Three stages of auditory processing are shown. Sound enters the cochlea and is transduced into a cochleagram (middle picture). A correlogram is then computed from the output of the cochlea by computing short-time autocorrelations of each cochlear channel. One frame of the resulting movie is shown in the bottom box. Reprinted with permission (Slaney & Lyon, 1993).

2.1 Cochlear Model

The cochlea converts pressure changes in the ear canal into neural firing rates that are transmitted through the auditory nerve. Sound pressure changes cause motion of the tympanic membrane which in turn transmits motion through the three ossicles (malleus, incus, and stapes) to the oval window of the cochlea. These vibrations are transmitted as waves of motion on the basilar membrane. The decreasing stiffness of the membrane, from base to apex, causes its mechanical response to change as a function of place. The net effect is that the basilar membrane acts like a set of band-pass filters whose center frequencies are place encoded as illustrated in Figure 4. Inner hair cells attached to the basilar membrane are bent by the movement, increasing the neural firing rate of the connected neurons. Since these hair cells only respond to motion in one direction, the signal is half-wave rectified. In addition the sensitivity and the characteristic impulse responses of the membrane vary as a function of the input level (Sachs and Young, 1979). Lyon's passive long wave cochlear model as implemented by Slaney (Slaney, 1988) is employed in this work to emulate the cochlea. Features of the model that are most important for this thesis are presented in the following subsections.

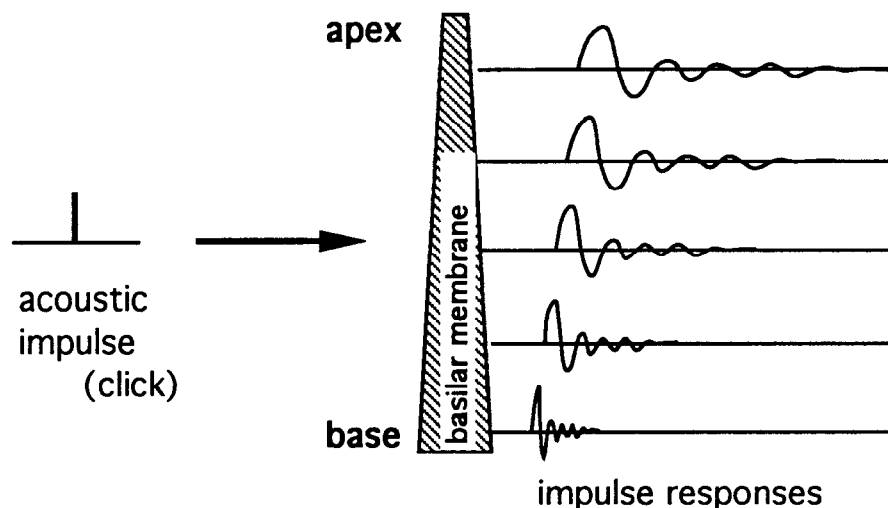


Figure 4. Impulse responses as a function of place on the basilar membrane. The center frequency and the bandwidth decrease logarithmically from base to apex. A time delay is introduced as the sound propagates from the base to the apex.

2.1.1 Cochlear Filter

A set of second-order sections is cascaded to model the behavior of the basilar membrane. This combination of filters is called the cochlear filter bank. The outputs of the early stages represent the response near the base of the cochlea, and later stages are closer to the apex. The center frequency of each channel decreases exponentially and then linearly with increasing channel number. Figure 5 shows a block diagram of the cochlear filter. The pre-emphasis stage is a pair of second-order filters which model the effect of the middle ear. The transfer function for each channel is represented by $H(\lambda_i, \omega)$ and the individual stages are represented by $F(\lambda, \omega)$, where λ_i is the i^{th} channel which represents location along the basilar membrane. Figure 6 shows the normalized transfer function magnitude for every tenth channel.

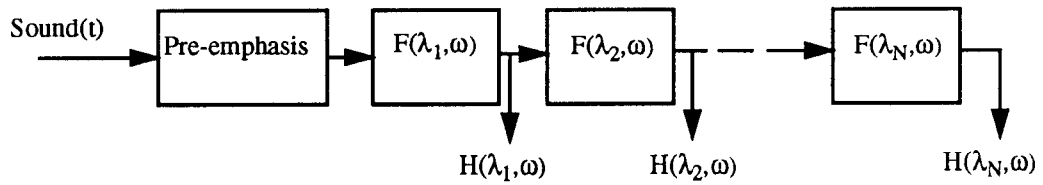


Figure 5. Cochlear filter bank diagram. The transfer function for each second-order stage is defined by F . The transfer function for a channel of the cochlear model is the product of all the preceding stages and is represented by H .

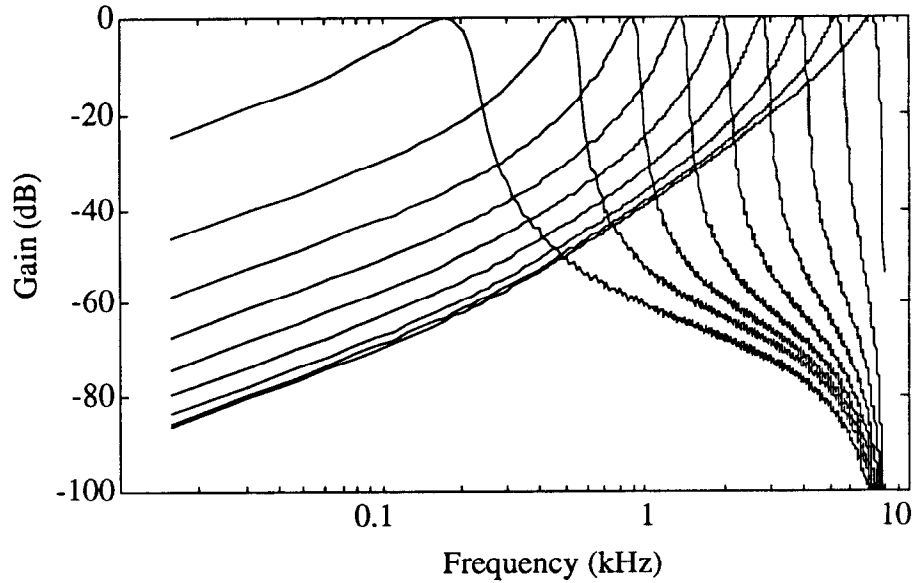


Figure 6. Normalized cochlear filter bank transfer functions ($H(\lambda_i, \omega)$) for every tenth channel.

2.1.2 Automatic Gain Control (AGC)

To accommodate the large range of sound pressure levels, the auditory system must automatically adjust its response. This adjustment is implemented in the cochlear model via a gain control at each channel that is varied as a function of the response; as the response increases in a given frequency region, the gain is subsequently reduced. The AGC in this model, however, does not affect the cochlear filter transfer functions as has been observed in the auditory system. The coupling that occurs between locations on the basilar membrane is also modelled in the AGC. The input to the AGC system is the half-wave rectified output of a channel of the cochlear filter bank. The AGC system multiplies the input by a level dependent gain. The AGC loop filter tries to maintain the average output at or below a target value, t , within some time constant, τ . The AGC for each channel is actually made up of a set of four AGC filters cascaded together. Each of the four filters has a different time constant varying from 640 to 10 ms. If there is sufficient power in the input signal, the state variable as depicted in Figure 7 can exceed unity which can cause the AGC filter to become unstable. This is avoided by limiting the state variable to a maximum value

of unity in the cochlear model (Slaney, 1988). A slight modification is included here to insure the gain is never zero by limiting the state variable to be less than unity by some a small amount. This modification is necessary to guarantee that the AGC is invertible.

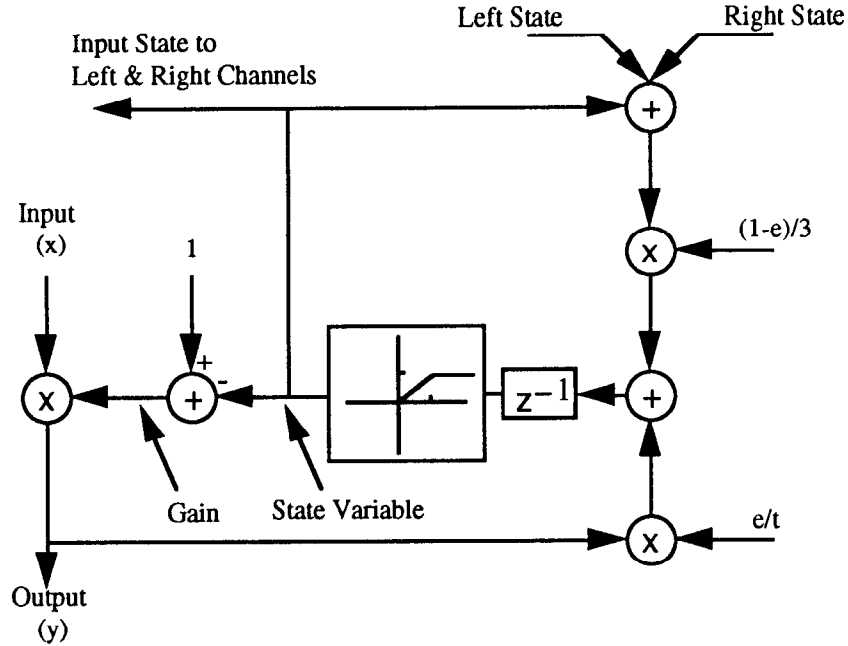


Figure 7. AGC block diagram. The parameter e ($e=T/\tau$, where T is the sampling period) sets the time constant for the filter and t sets the target output value. The left and right states allow coupling between channels. The state variable is limited to be less than unity so that the gain can never be zero, and is non-negative because the input x is half-wave rectified.

2.2 Correlogram Calculation

As described earlier, each frame of the correlogram is composed of a set of short-time autocorrelation functions for all the outputs of the cochlear model. The purpose of describing the autocorrelation function is two-fold. First, an understanding of the calculation is necessary in order to appreciate why the correlogram displays fine time structure, and second it is important to relate the correlogram to the Short Time Fourier Transform Magnitudes (STFTMs) upon which the correlogram inversion is based. Although it may not be immediately apparent, the short-time autocorrelation function can be estimated by the inverse Fourier transform of the squared magnitude of the STFT.

The autocorrelation function for a wide-sense stationary signal, $R_{xx}(\tau)$, is defined to be

$$R_{xx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t) x(t - \tau) dt \quad (1)$$

This definition states that the signal is shifted relative to itself, multiplied, and then averaged. If there is an inherent periodicity or pitch to the signal, the autocorrelation function will have a local maximum wherever the lag τ is equal to an integral multiple of the periodicity. Thus, the autocorrelation function reveals the periodicity in a signal.

This definition assumes that the signal is wide-sense stationary. For many signals, like speech, this assumption is not valid. An autocorrelation function that is evaluated for a short-time window (finite T) is called the short-time autocorrelation function. To evaluate the autocorrelation function over the entire signal, many such windows or frames of data must be analyzed. The original signal is windowed by multiplying by the time-domain function $w(t)$. The window is shifted to examine different portions of the signal. Thus, the short-time autocorrelation function is defined as

$$\hat{R}_{xx}(t, \tau) = \int_{-\infty}^{\infty} w(s - t) x(s) w(s - t - \tau) x(s - \tau) ds \quad (2)$$

The time, t , defines how far the window has been shifted for that frame of the autocorrelation functions.

To produce the correlogram, the short-time autocorrelation is calculated for each channel in the cochlear model. Such an operation can be viewed as applying a bank of autocorrelators to the outputs of the cochlear model, one autocorrelator for each channel. Since the autocorrelation is evaluated at discrete moments in time, there is one frame for each window position. The correlogram is produced when the frames are played back as a movie.

In actuality, all of this is implemented as operations on discrete signals. For a discrete

signal, $x(n)$, a window of data is defined to be

$$x_{w(s)}(n) = w(n-s) \cdot x(n) \quad (3)$$

$$\text{where } w(l) = 0 \text{ for } (l < 0) \text{ and } (l > L-1). \quad (4)$$

The estimate of the discrete short-time autocorrelation function is defined as

$$\hat{R}_{xx}(s, k) = \frac{1}{L} \sum_{n=s}^{s+L-|k|-1} x_{w(s)}(n) \cdot x_{w(s)}(n+|k|) \quad (5)$$

Although this computation can be performed in many ways, in Slaney's implementation it is computed using discrete Fourier transform techniques.

As discussed above, the autocorrelation function is related to the square of the magnitude of the Fourier transform of a signal. This is derived from the Khintchine-Wiener relations, which show that the power spectral density is equal to the Fourier transform of the autocorrelation of a signal:

$$S_x(\omega) = \int_{-\infty}^{\infty} R_{xx}(\tau) e^{-j\omega\tau} d\tau \quad (6)$$

Furthermore, the magnitude of the Fourier transform is the square root of the power spectral density. This relationship is also true for the short-time discrete case. Thus, knowing the correlogram or short-time autocorrelation function is equivalent to having the magnitude of the Fourier transform. By converting the autocorrelation functions to STFTMs, algorithms previously developed to estimate a signal based on its STFTMs can be used. These algorithms are discussed in detail in Section 4.

3 Sound Waveform Estimation from Cochleagrams

The cochlear band-pass filters, the half-wave rectification (HWR), and the automatic gain control (AGC) are combined in the processing performed by the cochlear model. In order to estimate the original time sequence of the signal that entered the cochlear model, these operations must be inverted. The following sections develop the necessary operations to estimate the sound waveform from its cochleagram. The first section presents a method to invert the cochlear filter, resynthesizing the sound from the output of the cochlear filter bank. The second section examines the iterative approach for inverting the half-wave rectification, and the third section presents the AGC inversion. Section 4 of this thesis will address the problem of converting the correlogram into a cochleagram which can be inverted using the methods described here.

3.1 Cochlear Filter Inversion

This section develops a method to reproduce the input to the cochlear model given an estimate of the cochlear filter bank outputs. If this estimate has no error, then the input can be reconstructed by filtering the estimate for any one channel by the reciprocal of the transfer function for that channel of the filter bank. This solution, although theoretically possible, produces poor results when the estimate of the cochlear filter bank outputs is in error.

The forward transfer function for any channel of the cochlear filter bank is a band-pass filter. The reciprocal of any of these filters has a very large gain outside the passband. If the estimate of the output of cochlear filter bank for a given channel contains energy outside the passband, then the large gain of the reciprocal filter will amplify the error, thus producing a poor reproduction of the input signal to the cochlear model. Motivated by the work of Yang et al. (1992), we employ a less common technique that reduces the amplification of the error.

The inversion of a linear filter bank must produce an estimate of the original input signal, $x(t)$, given the output of each channel, $y(\lambda, t)$. The output of each channel is defined as the

convolution of the input signal with the impulse response, $h(\lambda, t)$, or

$$y(\lambda, t) = \int_{-\infty}^t x(\tau) * h(\lambda, t - \tau) d\tau \quad (7)$$

An estimate of $x(t)$ can be realized by convolving the signal $y(\lambda, t)$ by using the matched filter $h(\lambda, -t)$. The estimate of $x(t)$, $\hat{x}(t)$, is given by the sum of convolutions

$$\hat{x}(t) = \sum_{\lambda} y(\lambda, t) * h(\lambda, -t) \quad (8)$$

Each term in the summation is the output of a channel convolved with the time-reversed impulse response of that channel. Substituting $-t$ for t in the last equation creates a time-reversed estimate of the input signal,

$$\hat{x}(-t) = \sum_{\lambda} y(\lambda, -t) * h(\lambda, t) \quad (9)$$

This subtle change is significant because it allows the inversion to be implemented with only minor modifications to the cochlear filterbank. However, it is necessary to time reverse both the output of the channels and the result to obtain the estimate of the original signal.

The estimate, $\hat{x}(t)$, is transformed to the Fourier domain to determine if it is an accurate representation of $x(t)$. Taking the Fourier transform of Equation 9 we obtain

$$\hat{X}(-\omega) = \sum_{\lambda} Y(\lambda, -\omega) H(\lambda, \omega) \quad (10)$$

Expansion of $Y(\lambda, -\omega)$ using Equation 7 creates

$$\hat{X}(-\omega) = \sum_{\lambda} X(-\omega) H(\lambda, -\omega) H(\lambda, \omega) \quad (11)$$

which, by noting that $H(\lambda, -\omega) = H^*(\lambda, \omega)$, can be simplified to

$$\hat{X}(\omega) = X(\omega) \sum_{\lambda} |H(\lambda, \omega)|^2 \quad (12)$$

This result clearly shows that $\hat{x}(t)$ is equal to $x(t)$ if the summation of the transfer functions across all the channels at each frequency equals one:

$$\sum_{\lambda} |H(\lambda, \omega)|^2 = 1 \quad (13)$$

Because each term of the summation, $\sum_{\lambda} |H(\lambda, \omega)|^2$, is real valued, Equation 12 also shows that the filtering operation defined by Equation 9 removes the phase changes produced by the cochlear filter even if Equation 13 is not satisfied.

The phase inversion is implemented by slightly modifying the cochlear filter structure shown in Figure 5 and running the signals in *backwards*. Specifically, the time-reversed output of each channel of the cochlear filter, $y(\lambda, -t)$, is used as input to the same stage of the inverse cochlear filter structure and summed with the output of the preceding stage, as shown in Figure 8. This approach allows all the channels to be inverted simultaneously. The transfer function for each stage and each channel are the same as in the cochlear filter, however the order of the stages is reversed. The output, $\hat{x}(-t)$, is then time-reversed to produce $\hat{x}(t)$.

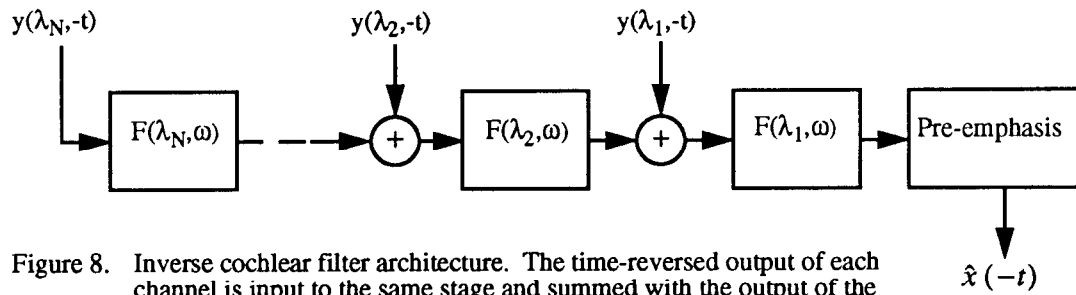


Figure 8. Inverse cochlear filter architecture. The time-reversed output of each channel is input to the same stage and summed with the output of the preceding stage. The order of the stages is the reverse of that for the cochlear filter.

If the spectral-tilt transfer function, $\sum_{\lambda} |H(\lambda, \omega)|^2$, is not equal to one, then the estimate, $\hat{x}(t)$, must be filtered by the reciprocal of the spectral tilt. Because of the nature of the

spectral tilt, shown in Figure 9, this operation is highly sensitive to noise at both low and high frequencies. To reduce the problem of having too much gain in these regions, the spectral tilt correction is limited so it never has a gain greater than 100 dB. This is an arbitrary choice which might be changed for different applications.

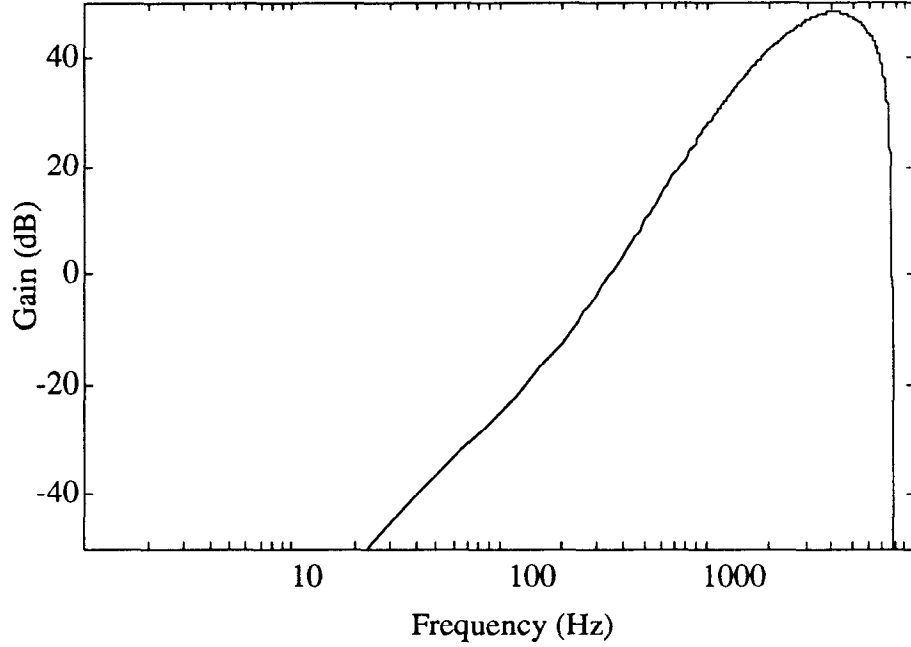


Figure 9. The spectral gain that results from the forward and backward processing through the cochlear filter. The tilt from low to high frequencies is termed the spectral tilt.

A spectral-tilt inversion function, $H_{st}^{-1}(\omega)$, is defined that satisfies this criterion,

$$H_{st}^{-1}(\omega) = \max\left(\frac{1}{\sum_{\lambda} |H(\lambda, \omega)|^2}, 10^5\right) \quad (14)$$

Noise present in the estimate, $\hat{x}(t)$, is modified by the spectral tilt inversion. In certain portions of the spectra the noise is attenuated or only slightly amplified. However, the spectral tilt inversion transfer function amplifies noise greatly at the lowest and highest

frequencies, and increases the sensitivity of the inversion to noise. Thus, the spectral-tilt inversion transfer function shown in Figure 10 can also be defined as the sensitivity function. When using the original cochleagrams this is not a significant problem, since there is not much noise. However, when one wants to estimate the signal from the half-wave rectified cochleagram signals or from the correlogram signals, this problem becomes acute. This issue is examined further in the next section which addresses half-wave rectification inversion.

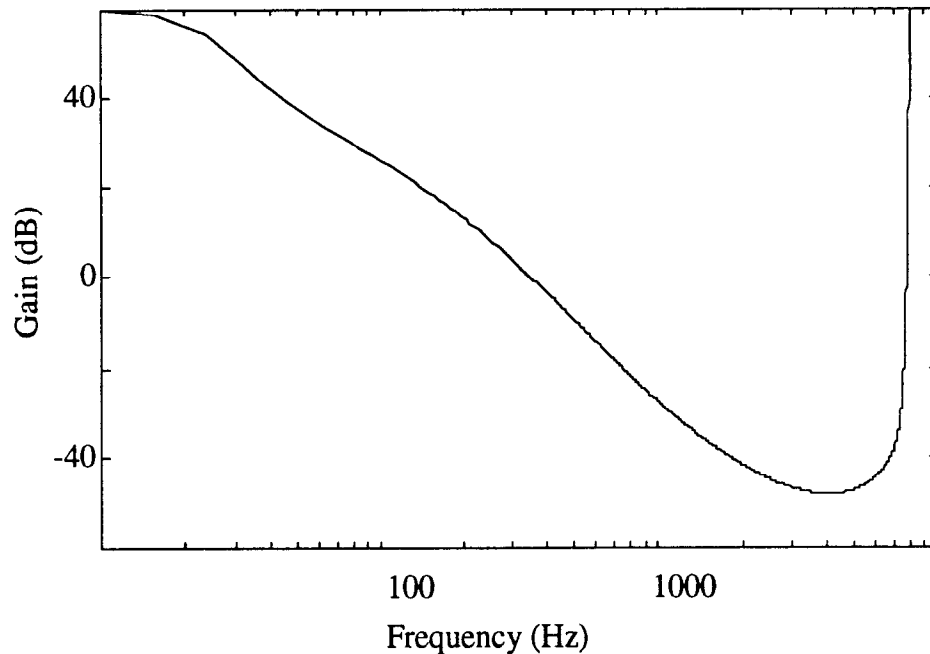


Figure 10. Spectral tilt inversion transfer function. Also defined to be the sensitivity function for the spectral tilt inversion. Noise with spectral content at frequencies where the spectral tilt inversion function is greater than one will be amplified.

If the computation of the cochleagram does not introduce large errors, the inversion of the cochlear filter bank can be performed with little error. Since the filtering is linear, the accuracy of the process can be measured by its overall impulse response. To determine the accuracy, we filtered a unit impulse through the cochlear filter bank, inverted it, and then compared the results to the original impulse. Clearly, no inversion error can be seen at this scale.

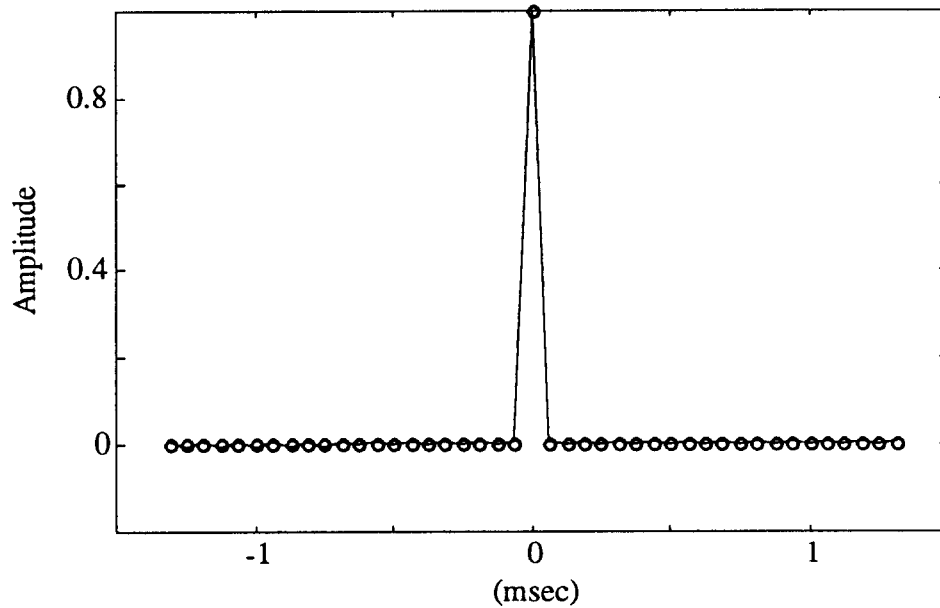


Figure 11. The solid line connects the samples of the impulse reconstructed from the outputs of the cochlear filter bank. The samples from the original impulse are shown as circles (o).

To more accurately quantify the error, the signal-to-error ratio (SER) power spectrum was calculated, and the results are displayed in Figure 12. The error is defined to be the difference between the original impulse and the estimate. Clearly, the error is extremely small in the range from about 200 Hz to 7 kHz.

To summarize, a method to invert the cochlear filterbank has been successfully developed and implemented. The present design of the spectral tilt inversion filter is highly sensitive to noise in the low and high frequencies. In the following section minor adjustments are included to the inverse filter when the half-wave rectification inversion is performed.

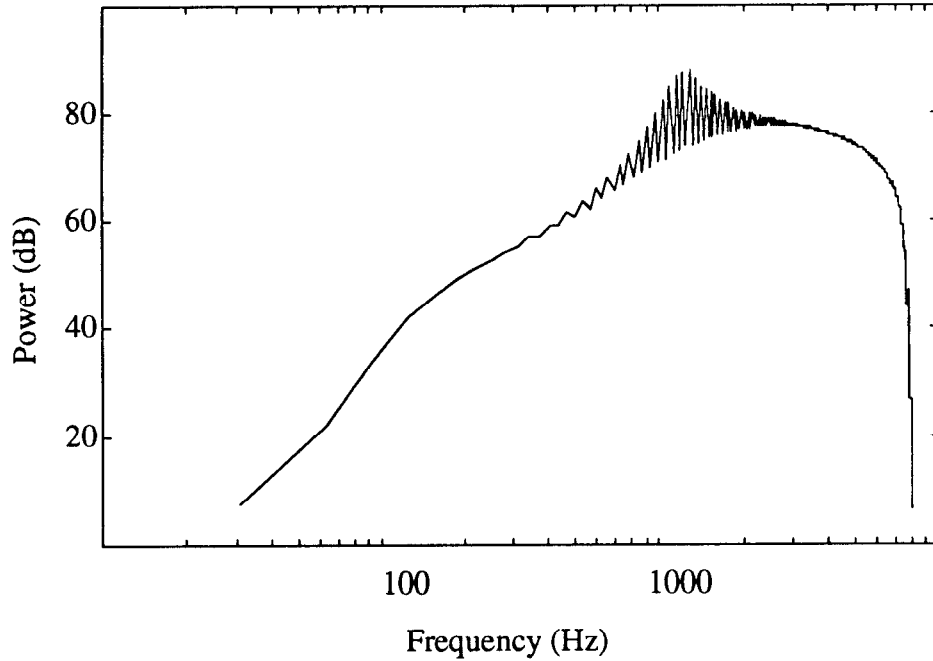


Figure 12. The SER of the impulse reconstructed from the outputs of the cochlear filter bank and the original. The error is the difference between the original and reconstructed impulses.

3.2 Inverting Half-Wave Rectification

Half-wave rectification of outputs of the cochlear filter bank can create distortion harmonics of the band-pass signal within each channel. If these band-pass signals occupy a narrow band, then the additional distortion products that result will fall outside this frequency region of interest. If the bands were sufficiently narrow, the half-wave rectification could be inverted by suppressing the frequency content outside the pass-band for each channel of the cochlear filter bank. Unfortunately, this requirement is not satisfied, and a more sophisticated technique must be used.

The process by which a band-pass signal can be estimated from its half-wave rectified representation will be developed first in a heuristic manner to understand the fundamental

approach. It will then be related to a general method of reconstructing signals from a partial representation known as the method of convex projections.

The following symbols appear in the development of this technique:

$x_{hwr}(t)$	The half-wave rectified representation of the original band-pass signal $x(t)$.
$\hat{x}(t)$	The estimate of $x(t)$. If the estimate is perfect, $\hat{x}(t) = x(t)$.
$B_x(\omega)$	A unity-gain band-pass filter with zero phase delay at all frequencies. $B_x(\omega)$ defines the portion of the spectrum that $x(t)$ occupies.

Given that a signal, $x(t)$, is band-limited to a region of the spectrum defined by $B_x(\omega)$, a method is developed to construct an estimate of $x(t)$ from its half-wave rectified representation $x_{hwr}(t)$. First the estimate $\hat{x}(t)$ is set equal to $x_{hwr}(t)$. Then the estimate is filtered by $B_x(\omega)$. After this operation the estimate is no longer guaranteed to have the same positive values as $x_{hwr}(t)$. In order to satisfy this restriction, the estimate is set equal to $x_{hwr}(t)$ wherever $x_{hwr}(t)$ is positive. The estimate must also be less than or equal to zero wherever $x_{hwr}(t)$ is equal to zero. Wherever the estimate exceeds zero in these portions, it is set to zero. These two time-domain adjustments set $\hat{x}(t)$ to have the correct positive values and to have nonpositive values where the signal is known to be less than or equal to zero. At this point, however, the estimate is no longer guaranteed to be band-limited. Thus, the filtering and correcting operations are repeated until some criterion is reached. The estimate is obtained after the positive values are fixed. This algorithm, illustrated in Figure 13, is used to estimate $x(t)$ given both $x_{hwr}(t)$ and $B_x(\omega)$. By examining these operations as projections onto closed convex sets, it can be shown that the algorithm converges to an optimal solution.

The fundamental idea of projection onto closed convex sets is to restrict the estimate to known properties of the signal being reconstructed. A set is defined to be convex if the midpoint of any two points belonging to the set is also a member of the set. That the signal to be estimated is a member of a convex set is important in that it allows a simple method to move to the closest signal (point) in the set given an estimate that does not belong to the set. The restriction that the set is closed is important for theoretical reasons, because it

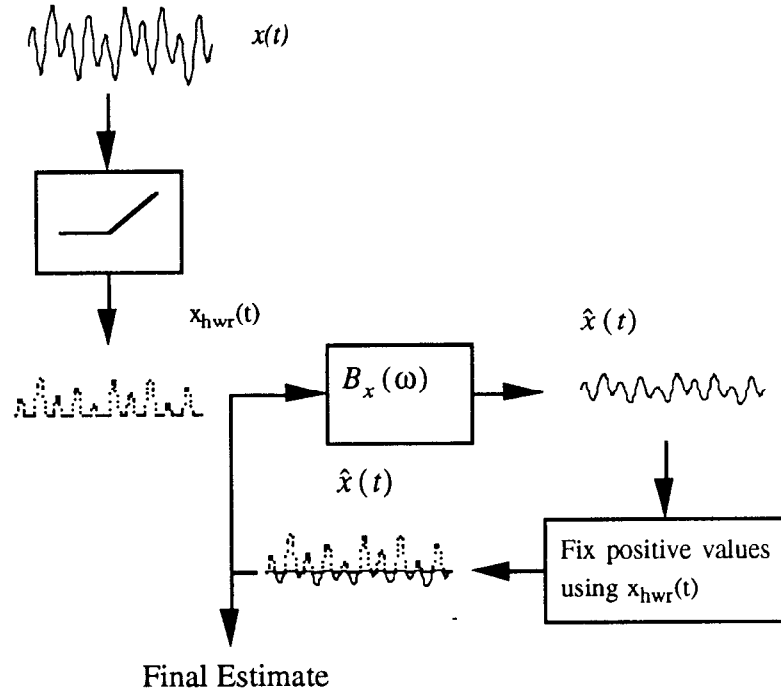


Figure 13. Algorithm to estimate a band-pass signal given its half-wave rectified representation.

guarantees that the limit of any number of iterations on a member of the set will also be in the set. For this discussion the signal is treated as a high dimensional vector. For the purpose of this algorithm, each known property of the original signal is used to restrict the estimate to a closed convex set. The intersection of these sets produces the best estimate of the signal. By definition the sets must intersect or else the signal to be estimated could not belong to both sets. Youla and Webb (1982) show that iteratively projecting onto each convex set to which the signal belongs will theoretically converge to a point in the intersection, and consequently produce an estimate of the original signal. Prior to applying this method, the signal that is to be reconstructed must have the convex sets to which it belongs defined and proven to be closed and convex, and the operators that project into these sets must also be defined. Although Yang et al. (1992) have shown that a band-limited signal and its half-wave rectified representation each create a closed convex set, the following non-rigorous proofs are provided.

The set defined by a half-wave rectified signal includes all signals that have those same positive values. Two signals within this set are chosen at random. Since they both have the same positive values their midpoint will also have the same positive values and will belong to the set, therefore it is convex. An operator that projects a signal into this set must simply set all positive values as defined by the half-wave rectified signal and insure that all other values of the signal are nonpositive. Thus, the information about the positive values of the original signal defines a convex set.

A band-pass signal is also contained within a convex set, but this set is defined in the frequency domain. Since all frequencies outside of the passband have zero energy for any signal in this set, the linear combination of any two signals cannot create energy outside the passband. Therefore, the midpoint of any two signals must remain within the set.

Projection into the set is implemented using an appropriate band-pass filter.

To implement the convex projection algorithm it is necessary to design the band-pass filter $B_x(\omega)$ for each channel of the cochlear filter bank. The transfer functions of the cochlear filter bank cannot be used, both because they do not have unity gain across the pass band, and because many of them are too broad. If $B_x(\omega)$ is too wide, the restriction in the frequency domain is not sufficiently stringent. On the other hand, if the filter is too narrow, the spectral content necessary to reconstruct the signal will be filtered out. Therefore, a finite impulse response filter (FIR) is designed based on the cochlear filter transfer functions. To design the filter it is necessary to determine what the passband is for each channel.

The transfer function of each channel is normalized to 0 dB at its maximum, and the passband for each channel is defined as that portion of the spectrum where the normalized transfer function exceeds some threshold. The threshold is determined by finding the value that minimizes the variance of the error between the output of the highest frequency channel and the estimate obtained by using the convex projection procedure. This is illustrated in Figure 14 with the highest frequency channel since it has the greatest bandwidth. The error as a function of the threshold is plotted. As Figure 14 shows, if the

threshold is too high or too low the error becomes large. From these results the threshold is chosen to be -15 dB; for each channel $B_x(\omega)$ is set to one at all frequencies where the cochlear filter has less than 15 dB attenuation.

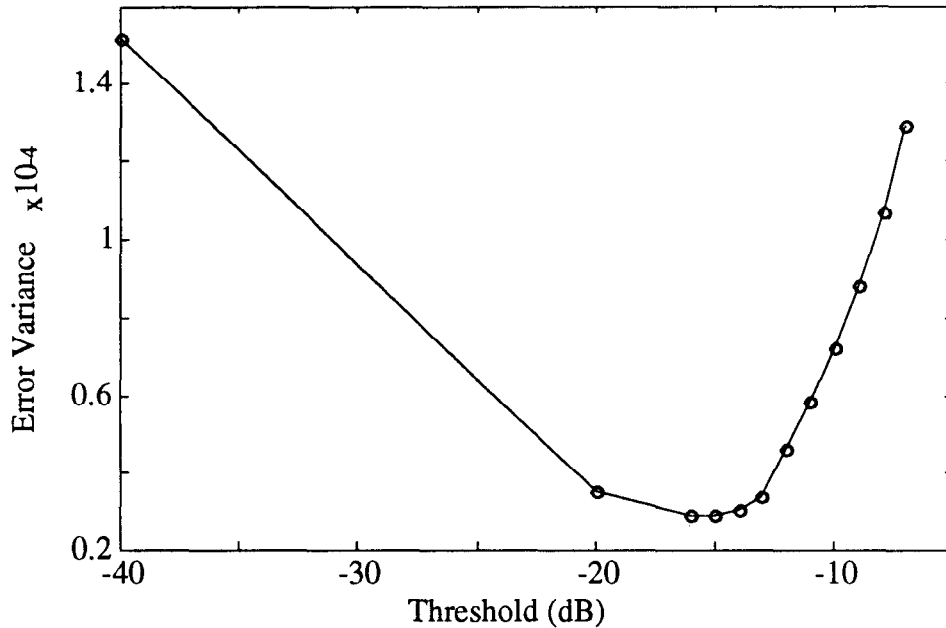


Figure 14. Variance of the error between the signal and its estimate for the highest frequency channel after 20 iterations. The error is a minimum when the threshold defining the passband is correctly set. The estimate is obtained by iteratively projecting onto convex sets defined by the positive values of the rectified signal and a band-pass filter defined by the threshold axis.

The criterion selected to halt the iteration process is the number of iterations. In order to assess what constitutes a reasonable number of iterations, a similar test is performed which evaluates the error as a function of the number of iterations. For this evaluation the threshold on the filter design is set at -15 dB.

At this point an examination of the spectral-tilt inversion developed in the preceding section is necessary. Although the inversion process converges, there is some error between the estimate and the original impulse. When the spectral-tilt filter inversion is implemented, it is known that the error can be amplified. In order to know both the reduction in the SER, and to determine if spectral inversion tilt needs to be modified, the

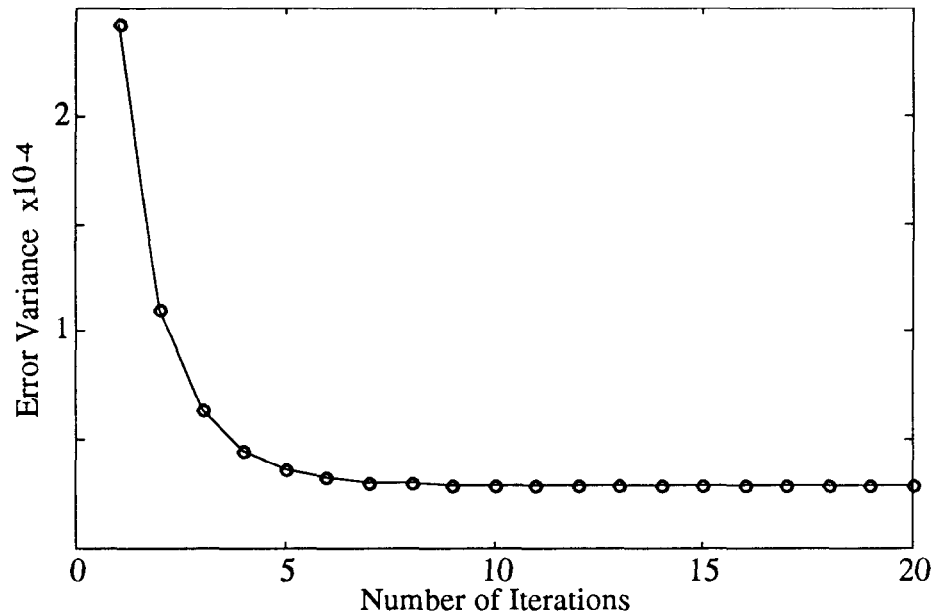


Figure 15. Variance of the error between the signal and its estimate as a function of the number of iterations. The estimate is obtained by iteratively projecting onto convex sets defined by the positive values of the rectified signal and a band-pass filter defined by the threshold axis. The threshold used to define the band-pass filter from the cochlear filter is set to -15 dB.

SER is evaluated as a function of the maximum relative gain allowed in the inversion filter. This maximum gain is the inverse of the attenuation threshold set in the previous section. Reducing the attenuation threshold used to design the spectral-tilt inversion filter produces an increased sensitivity to noise. However, increasing the threshold too much causes the resulting inverse filter to be inaccurate in the portion of the spectrum where significant energy is present. Therefore, the threshold must be set such that these two conflicting requirements are satisfied. This was done by increasing the threshold till the SER is reduced in the region where signal information is retained, approximately 200 Hz to 8 kHz. The results in Figure 16 show that as the threshold is reduced from -100 dB to -40 dB there is a dramatic increase in the SER near 200 Hz when the threshold is changed from -60 to -40 dB. Therefore, the threshold for designing the spectral-tilt inversion filter is set to -60 dB. Comparison of this result with that shown in Figure 12, indicates that inversion of the half-wave rectifier does create a noticeable reduction in the SER. These results may be

improved by redesigning the band-pass filters needed for the convex projection. However, it must be noted that to the author there is no perceptual difference between the original sound and the estimated sound.

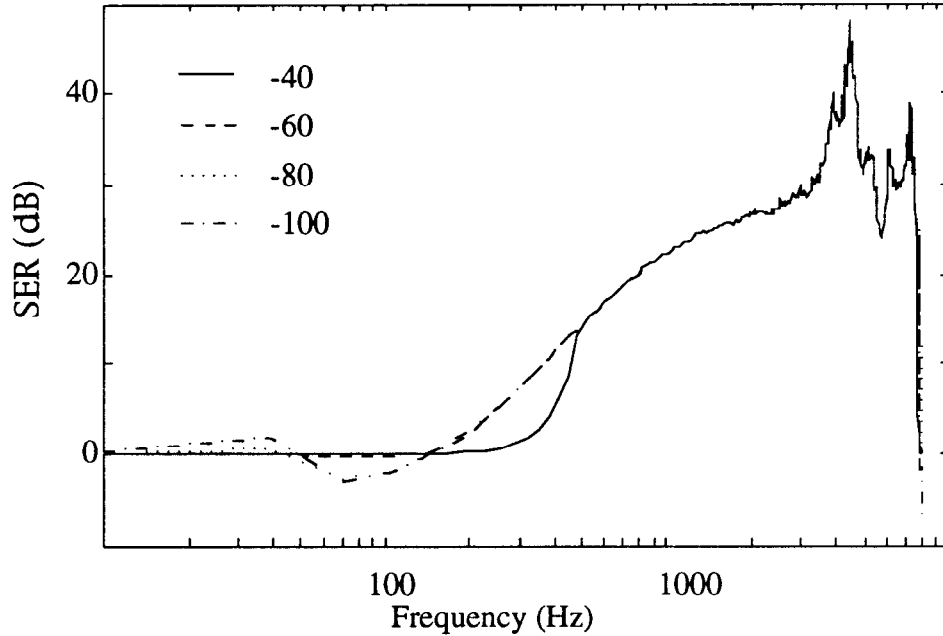


Figure 16. SER of impulse and estimated impulse. Frequency projection performed using 1024 point FIR filters. Spectral tilt inversion filter design is checked by varying the threshold from -100 to -40 dB.

The time-based comparison of the original impulse and the estimate created by this process, displayed in Figure 17, indicates the insignificance of the degradation.

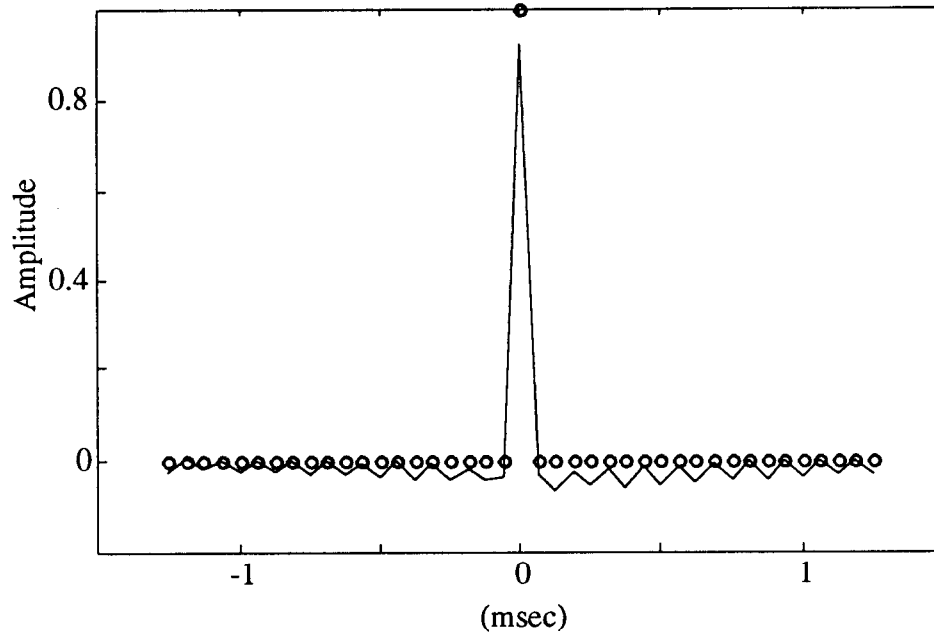


Figure 17. Original impulse (o) and the estimate obtained from the half-wave rectified cochleagram using FIR filter to perform the projection onto the convex set. Spectral inverse filter designed using -60 dB threshold.

Unfortunately, the processing time required to implement the FIR filters for the convex projections is very expensive, so an infinite impulse response (IIR) filter is investigated. Using the Butterworth filter approximation, the coefficients are developed using the threshold described above. Although the IIR filter requires less computation, it does produce phase distortion. In order to assess the effect of this distortion on the reconstruction of the impulse the SER plot and the reconstructed impulse are shown in Figure 18 and Figure 19.

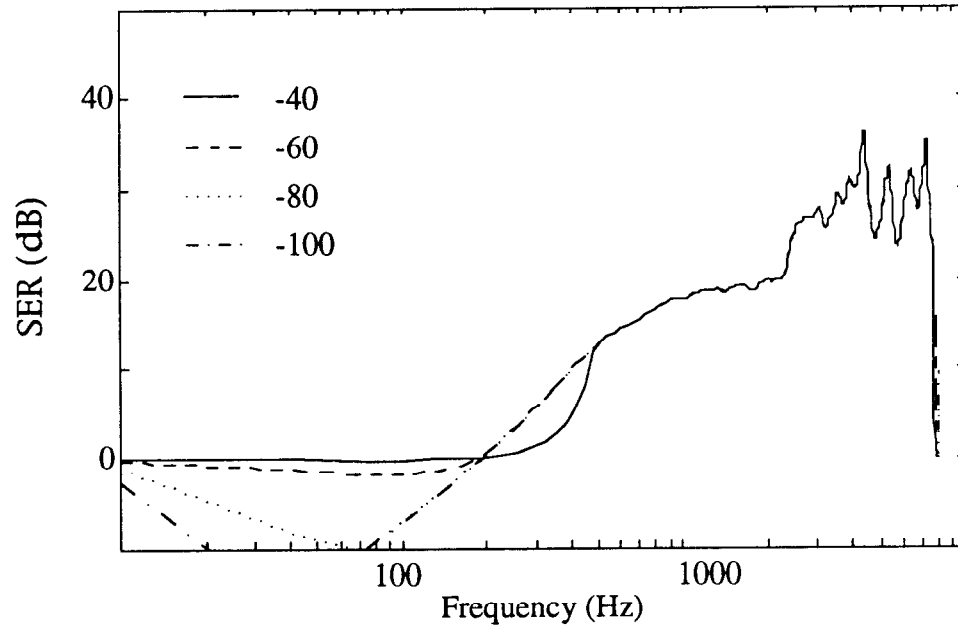


Figure 18. SER of impulse and estimated impulse. Frequency projection performed using Butterworth filters. Spectral tilt inversion filter design is checked by varying the threshold from -100 to -40 dB.

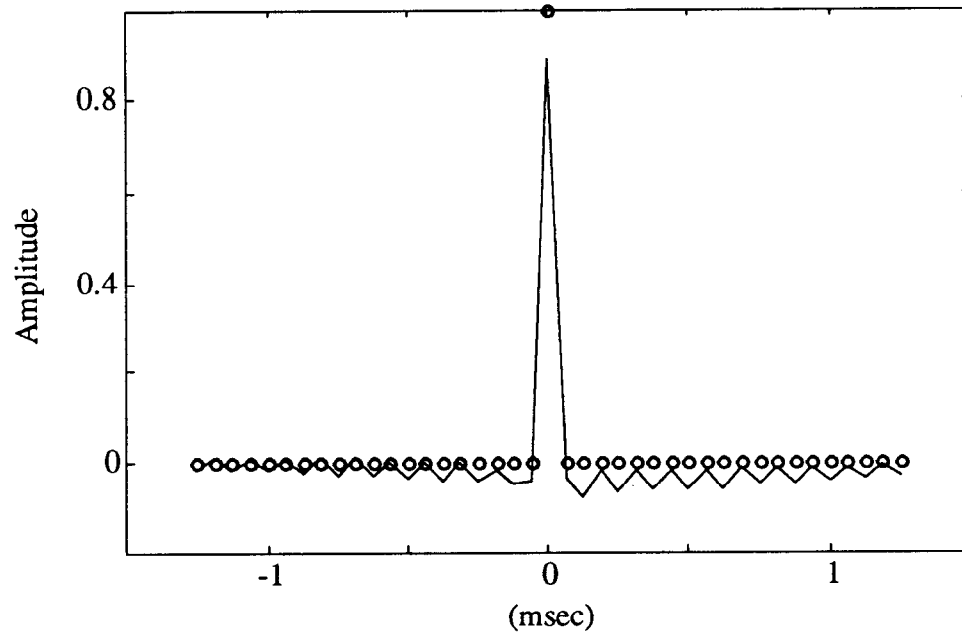


Figure 19. Original impulse (o) and the estimate obtained from the half-wave rectified cochleagram using IIR filter to perform projection onto convex set. Spectral inverse filter designed using -60 dB threshold.

Although the SER is reduced when using the Butterworth filter implementation, the perceptual degradation is not perceptible to the author. The Butterworth filter implementation is used to reduce the computational load. However, the experimental implementation of the inversion process developed in Matlab allows any arbitrary filter coefficients to be defined for the band-pass filters in the half-wave rectification algorithm.

3.3 Inverting Automatic Gain Control

Each channel in the cochleagram is scaled by a time varying function calculated by the AGC filter. In order to invert this operation, it is necessary to determine the scaling function at each instant in time. Upon examination of Figure 7 it is evident that the loop gain is dependent only on the AGC output, which is known. Thus, by swapping the input and output points, and dividing instead of multiplying by the loop gain, the AGC is inverted. The restructured filter to perform the inversion is shown in Figure 20. The AGC for each channel consists of four stages, therefore the AGC inversion also requires four stages processed in reverse.

The input to the AGC inversion will be the cochleagram estimated by the correlogram inversion algorithm. The error in cochleagram estimation will propagate into the AGC inversion. The sensitivity of the AGC inversion to this error is a function of both the gain and the error level. Therefore, it is also dependent on the input average power level and the target value selected.

The following variables are used in the derivation of the sensitivity:

\hat{x}	The estimate of the input.
N	The noise or error introduced. See Figure 20.
\hat{G}	The estimate of the gain, G .
\hat{S}	The estimate of the State Variable, S .
H	The transfer function of the low-pass filter in the AGC.

The result to be derived is a relation showing the dependence of \hat{x} on the error or noise, N , and on the input power level. Assuming that the gain function, G , is known, then the true

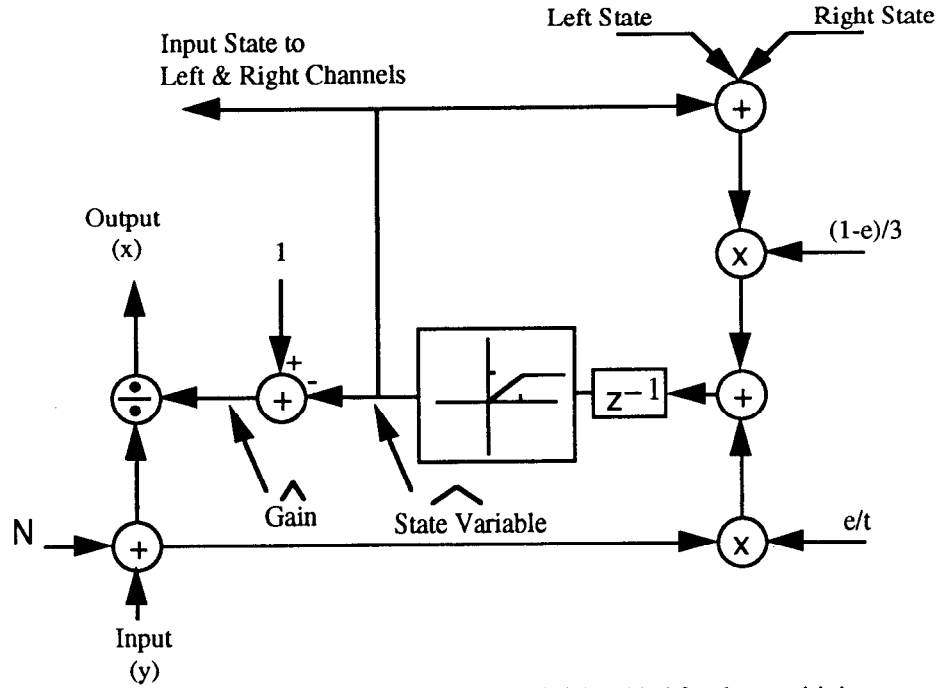


Figure 20. AGC inversion block diagram. Noise (N) is added for the sensitivity analysis. Both the gain and state variable have a hat indicating that they are estimates.

output of the AGC is defined to be

$$y = Gx \quad (15)$$

Furthermore, the estimate \hat{x} is related to y by

$$\hat{x} = \frac{y + N}{\hat{G}} \quad (16)$$

Substituting the relation for y results in

$$\hat{x} = \frac{\hat{G}x + N}{\hat{G}} = \frac{G}{\hat{G}}x + \frac{N}{\hat{G}} \quad (17)$$

which demonstrates that the error produces two effects. First it creates an additive error that is scaled by the estimated gain. Therefore, even if the gain is estimated exactly but is very small, the noise is amplified, causing a large error in the estimate of x . The second effect is multiplicative. The first element in Equation 17, $\frac{G}{\hat{G}}x$, causes the resulting estimate

to be scaled by $\frac{G}{\hat{G}}$ if the gain and the gain estimate are not equal. If G is small, then an error in \hat{G} translates into a large multiplicative error. So a small gain is likely to make a large source of error. Thus, both these errors are manifested when the gain is small. In general, inverting a small gain is obviously more difficult, since small errors in the gain result in large multiplicative errors which magnify the noise.

From the preceding analysis it is apparent that the gain must be maintained at a reasonable level to avoid problems in the AGC inversion. The gain is strictly a function of the input level. If the average input level is approximately twice the target value, then the gain will be approximately one-half. As the input level increases, the gain must be reduced. Thus, high input levels cause low gain, creating an inaccurate inversion.

The noise level can also cause the estimate of the gain to become even smaller. The estimated gain is defined to be

$$\hat{G} = 1 - \hat{S} = 1 - (S + H(N)) \quad (18)$$

where $H(N)$ defines a low-pass filter operation on the noise. If N has zero mean, then the low-pass filter will not add any energy to the estimate of G . However, if the noise has a bias there will be a corresponding error in the gain estimate. A cochleagram estimated from the correlogram can be viewed as the true cochleagram plus some zero mean noise component. However, in the estimation algorithm it is known that the cochleagram is non-negative, therefore this estimate is half-wave rectified. This half-wave rectification can only cause the noise component to attain a positive bias. Equation 18 shows that a positive bias causes the estimated gain to be reduced. This effect is significant when the gain is small.

4 Correlogram Inversion

The previous section addresses the issue of estimating a sound signal from its cochleagram. The more difficult issue, which this section addresses, is the development and implementation of a methodology to estimate a cochleagram by inverting its correlogram.

Each channel in the cochleagram can be viewed as an independent time sequence. If a single channel can be reconstructed from its respective row in the correlogram, in theory the entire cochleagram can be reconstructed. As described in Section 2.2, each row in a correlogram is the time (lag) domain representation of the Short Time Fourier Transform Magnitude (STFTM) of the cochleagram. An algorithm developed by Griffin and Lim to estimate a time signal based on its STFTMs provides a suitable approach to resolve this issue (Griffin and Lim, 1984). This method is subtly altered to include a priori information about the signal and its relation to neighboring channels. The method described by Griffin and Lim is reviewed and then, motivated by the work of Roucos and Wilgus (1985), it is modified in order to improve the initial estimate. The fact that the cochleagram is half-wave rectified is used to improve the algorithm. Finally, redundant information between channels is incorporated into the algorithm.

Section 4.1 of this chapter describes how a single channel of the correlogram can be inverted by implementing the algorithm described by Griffin and Lim. A scheme to choose a better starting phase to improve the convergence rate of the Griffin and Lim algorithm is shown in Section 4.2. Finally Section 4.3 describes how information from adjacent channels can be used to further improve the estimate and reduce the number of iterations.

4.1 Signal Estimation from its Short Time Fourier Transform Magnitude

Let $x(n)$ and $X_w(mS, \omega)$ denote a real sequence and its STFT. The analysis window used to calculate the STFT, $w(n)$, is defined to be real and non-zero for $0 \leq n \leq L-1$. Applying the window to the sequence creates,

$$x_w(mS, l) = x(l) \cdot w(mS - l) \quad (19)$$

a windowed portion of the sequence. The variable S sets the amount of shift between windows and the index, m , is the window number. For each sequence of data so defined, the STFT is calculated to be

$$X_w(mS, \omega) = \sum_{l=-\infty}^{\infty} x_w(mS, l) e^{-j\omega l} \quad (20)$$

The STFTs created from a signal are unique and consistent, so that given the complete set of STFTs the signal can be reconstructed exactly. However, an arbitrary set of STFTs are not guaranteed to reconstruct a signal from which the STFTs can be recalculated. What is the best signal that can be estimated given such a set of STFTs, $Y_w(mS, \omega)$? Griffin and Lim (1984) were presented with this identical problem when trying to time compress a speech signal using its STFTs. They defined the best estimate to be a signal $x(n)$, such that the distance between $Y_w(mS, \omega)$ and $X_w(mS, \omega)$, defined to be

$$D = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} |X_w(mS, \omega) - Y_w(mS, \omega)|^2 d\omega \quad (21)$$

is minimized. Using Parseval's theorem the distance metric can be written as,

$$D = \sum_{m=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} [x_w(mS, l) - y_w(mS, l)]^2 \quad (22)$$

where $y_w(mS, l)$ is the inverse Fourier transform of $Y_w(mS, \omega)$. By finding the derivative of Equation 22 with respect to $x(n)$, setting it equal to zero, and solving for $x(n)$, the best signal can be determined. Taking the derivative leads to

$$D'' = \sum_{m=-\infty}^{\infty} 2 [x(n) w(mS - n) - y_w(mS, l)] w(mS - n) \quad (23)$$

and setting D'' to zero and solving for $x(n)$ produces

$$0 = x(n) \sum_{m=-\infty}^{\infty} w^2(mS-n) - \sum_{m=-\infty}^{\infty} y_w(mS, l) w(mS-n) \quad (24)$$

$$x(n) = \frac{\sum_{m=-\infty}^{\infty} y_w(mS, l) w(mS-n)}{\sum_{m=-\infty}^{\infty} w^2(mS-n)} \quad (25)$$

This result, termed the Least Square Error Estimation from the Short Time Fourier Transform (LSEE-STFT), states that the best estimate is obtained by overlapping and adding the windowed time series obtained from the STFT (Griffin and Lim, 1984).

To reduce computational and memory requirements, the analysis window can be designed so that the denominator in the previous equation is equal to unity. A rectangular window satisfying this condition is

$$w_r(n) = \begin{cases} \frac{\sqrt{S}}{\sqrt{L}}, & \text{for } (0 \leq n < L) \\ 0, & \text{otherwise} \end{cases} \quad (26)$$

$$\sum_{m=-\infty}^{\infty} w_r(mS-n)^2 = 1 \quad (27)$$

Furthermore, if the window length (L) is restricted to be a multiple of four times the window shift (S), then Griffin and Lim show that the sinusoidal window defined below also satisfies the requirement that the denominator of Equation 25 is equal to 1.

$$w_s(n) = \frac{2w_r(n)}{\sqrt{4a^2 + 2b^2}} \left[a + b \cos \left(2\pi \frac{n}{L} + \phi \right) \right] \quad (28)$$

By setting $\phi = \frac{\pi}{L}$ the window is made symmetric, so that $w(n) = w(L-n-1)$. Setting $a = 0.54$ and $b = -0.46$, results in a modified Hamming window, with the subtle

difference that the period of the modified sine wave is L instead of $L-1$. This sinusoidal window is used to produce the results in this thesis.

The signal estimation problem using a row of the correlogram, however, starts with only the STFTM and does not contain any phase information. Therefore, an approach using only the STFTM, $|Y_w(mS, \omega)|$, must be developed. Based on the derivation above, an iterative algorithm was developed by Griffin and Lim. An initial guess is made for the signal, $\hat{x}(n)$, and the STFT, $\hat{X}_w(mS, \omega)$, is calculated. The magnitude of this STFT is then replaced with the known magnitude $|Y_w(mS, \omega)|$ and this new STFT is used to recalculate the estimate, $\hat{x}(n)$, based on the result of Equation 25. The phase information for each STFT is calculated from the most recent estimate of the signal, while the magnitude is always set back to that which was originally supplied. Since the magnitude is employed in this method it is termed the Least Squares Error Estimation using the Short Time Fourier Transform Magnitude (LSEE-STFTM).

Another distance measure

$$\hat{D} = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} [|X_w(mS, \omega)| - |Y_w(mS, \omega)|]^2 d\omega \quad (29)$$

has been shown by Griffin and Lim to decrease monotonically and to converge to a set of critical points as the iteration progresses. This algorithm, however, is not guaranteed to converge to a global minimum.

The distance measure defined in Equation 29 can vary dramatically depending on the power of the signal. By summing the integrated power spectra over all the windows and dividing by the distance metric the signal-to-error ratio is estimated to be

$$SER = \frac{\sum_{m=-\infty}^{\infty} \int_{-\pi}^{\pi} |Y_w(mS, \omega)|^2 d\omega}{\sum_{m=-\infty}^{\infty} \int_{-\pi}^{\pi} [|X_w(mS, \omega)| - |Y_w(mS, \omega)|]^2 d\omega} \quad (30)$$

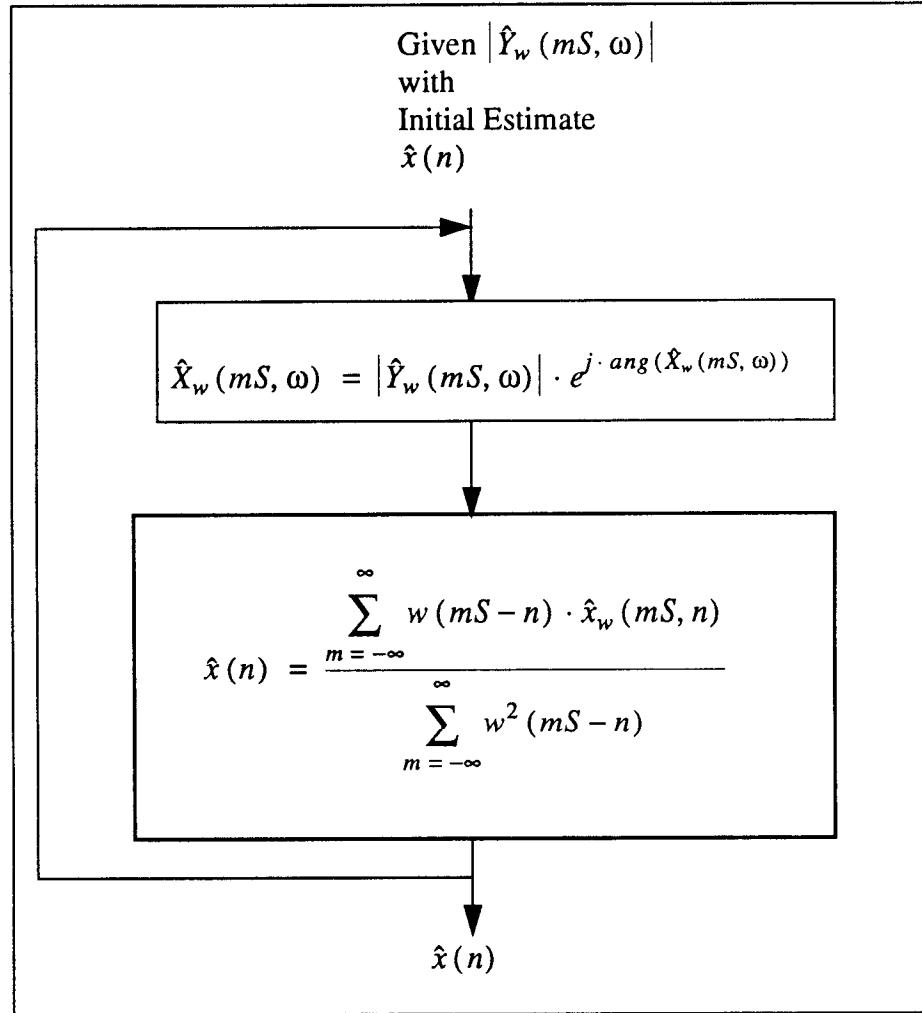


Figure 21. LSEE-STFTM algorithm

This section has reviewed a method developed by Griffin and Lim to estimate a signal based on its STFTM. With this approach, each channel in the cochleagram can be estimated from its respective row in the correlogram. In order to increase the convergence rate and to improve phase coherency between channels, the method is modified as described in the following section.

4.2 Improving the Initial Guess

The initial guess for $\hat{x}(n)$ profoundly affects the convergence rate and the result of the

LSEE-STFTM algorithm. Using a straight-forward procedure Roucos and Wilgus (1985) improved the initial estimate for performing time scale modification of speech. Time scale modification is used to change how long the signal is without changing other properties of the signal. A common example is speeding up or slowing down recorded speech while maintaining the original periodicity in the signal so that the pitch does not change.

Instead of overlapping and adding the windows assuming zero phase Roucos and Wilgus synchronize the most recent window of the original signal with the rest of the estimated signal. The amount that the m^{th} window is shifted, before being added to the present estimate, is obtained by maximizing the cross-correlation of the data in the m^{th} window of the original waveform to the estimated waveform up to the $m-1^{\text{st}}$ window. It is important to note that Roucos and Wilgus retained the phase of the window of data from the original waveform and then applied a linear phase shift. Thus, the internal structure of each window of data is retained. When inverting the correlogram, however, the internal phase structure of each window has already been set to zero. However, since speech is usually periodic in nature the same approach is taken, but the internal phase structure of each window has been set to zero. Therefore it is not anticipated that the improvement in the convergence rate will be as great as that achieved by Roucos and Wilgus.

By shifting the data windows by an amount, k , the denominator in Equation 25 becomes

$$c(n) = \sum_{m=-\infty}^{\infty} w^2(k + mS - n) \quad (31)$$

and is no longer guaranteed to be equal to unity. Therefore, $c(n)$ must be calculated in the algorithm described below.

The approach taken here calculates the shift using the window of data calculated from the STFTM, $|Y_w(mS, \omega)|$. From $|Y_w(mS, \omega)|$ the window of data is calculated by taking its inverse Fourier transform, resulting in

$$y_w(mS, n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |Y(mS, n)| e^{j\omega n} e^{-j\omega \frac{L}{2}} d\omega. \quad (32)$$

Careful inspection of this definition for $y_w(mS, n)$ reveals that an additional lag of $L/2$ (180 degrees) is included to center the signal in the analysis window defined in Section 4.1.

The first step of this algorithm is to initialize the estimate and $c(n)$ to

$$\hat{x}_0(n) = w(n) \cdot y_w(0, n) \quad (33)$$

$$c_0(n) = w^2(n) \quad (34)$$

The next window to overlap and add is shifted by the amount, k , that maximizes the cross-correlation

$$R_{xy_w}(k) = \sum_{n=mS-k}^{mS-k+L} \hat{x}_{m-1}(n) \cdot y_w(mS, n+k) \quad (35)$$

The magnitude of the shift is limited to one quarter of the window length. Once k_{\max} is found, it is used to overlap and add the m^{th} window in the following manner

$$\hat{x}_m(n) = \hat{x}_{m-1}(n) + w(mS, k_{\max} - n) \cdot y_w(mS, n + k_{\max}) \quad (36)$$

and the function $c(n)$ is updated to be

$$c_m(n) = c_{m-1}(n) + w^2(mS, k_{\max} - n) \quad (37)$$

This process is repeated until all the windows have been added to the estimate and then $\hat{x}(n)$ is divided by $c(n)$. The result of this process is used as the initial guess for the algorithm developed in the preceding section. Adding a shift or time delay to each window of data is equivalent to starting the Griffin iteration with linear non-zero phase. The iterative algorithm shown in Figure 21 assumes an initial phase that is zero for each frequency in each window. A non-zero initial phase, as determined by Equation 35, supplies an initial estimate that is expected to have a STFTM that is closer to the signal to

be estimated.

To determine if this adjustment of the initial estimate improves our ability to estimate a signal from a set of STFTMs, a 300-Hz sinusoid modulated at 60 Hz is reconstructed from its STFTMs. Figure 22 shows that the initial error is approximately half that obtained without including the synchronization. This example also shows that, as expected, the error is continually reduced as the number of iterations is increased. Furthermore, the error is always smaller for the same number of iterations when the windows are synchronized. After approximately five iterations, however, the error is not reduced significantly by more iterations.

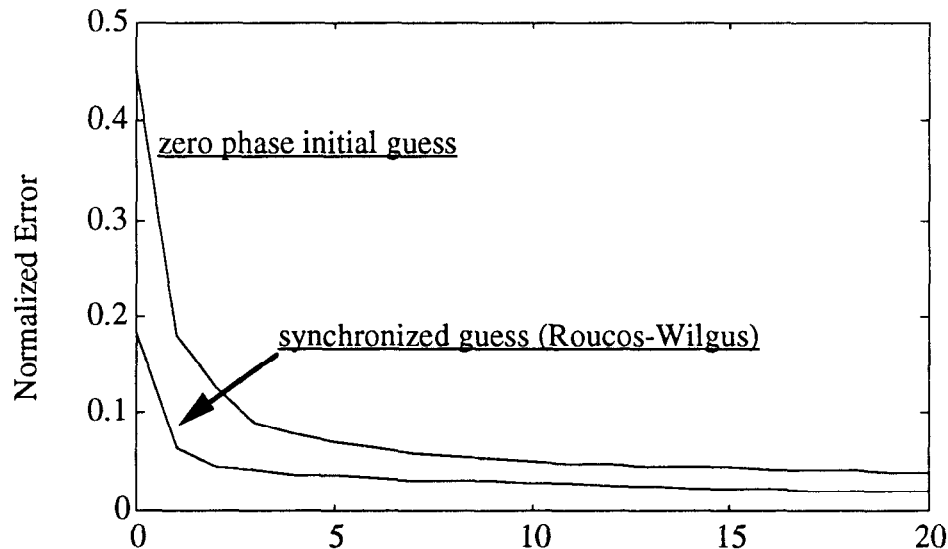


Figure 22. Normalized error as a function of the number of iterations for two techniques. The first assumes no phase information in the initial guess and sets the phase to zero. The second uses a synchronized overlap technique to produce an initial phase estimate.

4.3 Including Knowledge of Signal Characteristics

The reconstruction of the cochleagram, to this point, has only dealt with the STFTM of a single arbitrary signal. Information that is known a priori can be exploited to reduce the computational load and to create a better estimate. Two basic features concerning the original signals are known:

The signals are linearly related prior to being half-wave rectified and scaled by the gain control.

The signals are half-wave rectified.

The half-wave rectification restriction can be included in the reconstruction by half-wave rectifying the signal after each iteration of the overlap-and-add procedure. This operation is a convex projection as defined in Section 3.2. The estimate is half-wave rectified after every step of the iteration in the overlap-and-add procedure.

Although the linear relationship between channels is a heuristically appealing constraint, it is not guaranteed to improve the inversion process. This is because the signals have been non-linearly filtered by the AGC and the half-wave rectifier. Motivated by the correspondence between channels, this section describes an attempt made to predict the phase of one channel from that of a preceding one. Two methods of doing this are included: the first uses the phase of the STFTs of the preceding channel to estimate the following channel's phase, and the second uses the previous channel's phase and also includes the phase change predicted by the cochlear filter. These two methods are termed phase prediction and delta phase prediction, respectively. No matter which method is used, the synchronized overlap-and-add algorithm is employed to invert the first channel. The phase estimation procedure is only helpful for providing a better initial guess for estimating each channel.

Phase prediction and delta phase prediction are fundamentally the same algorithm. Therefore, they are described together here and the additional step in the delta phase method is identified. In the following description it is assumed that the signals for all the channels up through and including λ_1 have been estimated. From the signal estimated for channel λ_1 , $x(\lambda_1, n)$, a set of STFTs, $X_w(\lambda_1, mS, \omega)$, are calculated and the phase information is retained. The predicted phase for each window of the next channel, λ_2 , is stored in the quantity

$$\angle \hat{X}_w(\lambda_2, mS, \omega) \equiv \frac{X_w(\lambda_1, mS, \omega)}{|X_w(\lambda_1, mS, \omega)|} \quad (38)$$

where the \angle symbol indicates a unity gain complex vector containing only phase information.

If the delta phase prediction is being employed, at this juncture the anticipated phase change between channel λ_1 and λ_2 is included. It is not assumed that the two channels are adjacent, therefore the phase change across a number of stages in the cochlear filter is included. The phase estimate is changed to

$$\angle \hat{X}_w(\lambda_2, mS, \omega) \equiv \angle \hat{X}_w(\lambda_2, mS, \omega) \cdot \frac{F_{\lambda_1+1}}{|F_{\lambda_1+1}|} \cdot \frac{F_{\lambda_1+2}}{|F_{\lambda_1+2}|} \cdots \frac{F_{\lambda_2}}{|F_{\lambda_2}|} \quad (39)$$

The STFTMs and their estimated phase functions are combined to create a set of estimated STFTs

$$\hat{X}_w(\lambda_2, mS, \omega) = Y_w(\lambda_2, mS, \omega) \angle \hat{X}_w(\lambda_2, mS, \omega) \quad (40)$$

which are used to create the windows of data

$$\hat{x}_w(\lambda_2, mS, \omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}_w(\lambda_2, mS, \omega) e^{j\omega n} d\omega \quad (41)$$

Finally these sequences are combined in the overlap-and-add method to create the initial estimate of the signal for channel λ_2 ,

$$\hat{x}(\lambda_2, n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n) \cdot \hat{x}_w(mS, n)}{\sum_{m=-\infty}^{\infty} w^2(mS-n)} \quad (42)$$

which is used to initialize the Griffin-Lim inversion scheme described in Section 4.1.

Because these two approaches predict better estimates of the phase it is anticipated that they will reduce the error calculated when compared to the sequential overlap-and-add method. Furthermore, it is assumed that the delta phase prediction technique will perform the best. To compare the different techniques, an impulse is entered into the cochlear model and the resulting correlogram is calculated. The normalized error, as defined in Equation 30, is calculated for each channel after every iteration. Figure 23 shows the error for all channels without any iterations for each of the three methods: sequential overlap and add, phase prediction, and delta phase prediction. The error for the first channel is the same for all three, since it is always estimated using sequential overlap and add. Clearly when the phase information is included the error drops significantly. However, the difference between the performance of the phase and delta phase prediction techniques is not consistent.

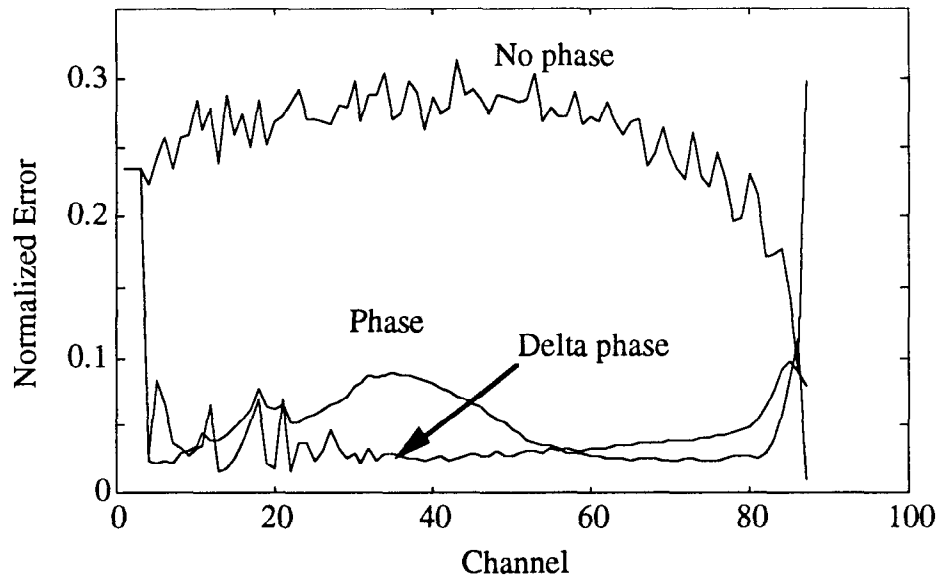


Figure 23. Correlation inversion error with no iterations as a function of channel number and three initial phase estimates. The error for three techniques is compared: overlap and add, phase prediction and delta phase prediction. An impulse is used as input to the cochlear model.

The results are presented after ten iterations in Figure 24. Here it becomes obvious that the

delta phase technique is performing best across most of the channels. It is also apparent that the sequential overlap-and-add approach has improved dramatically in the low-frequency channels (frequency decreases logarithmically with increasing channel number).

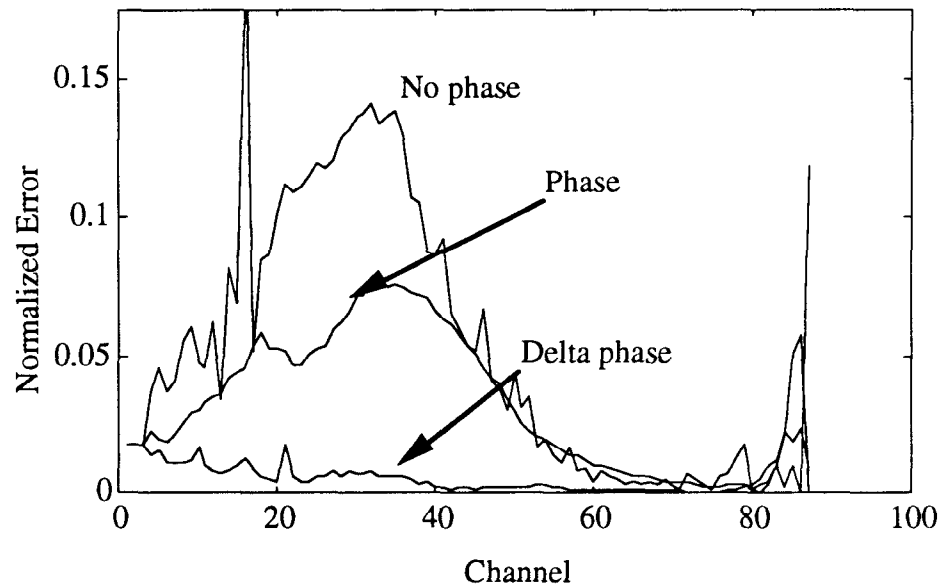


Figure 24. Correlation inversion error after ten iterations as a function of channel number and three initial phase estimates. The error for three techniques is compared: overlap and add, phase prediction and delta phase prediction. An impulse is used as input to the cochlear model.

Although it appears that the delta phase prediction method performs best, it is difficult to extrapolate the results from an impulse to a complex signal such as speech. However, it is still interesting to further examine the error of the delta phase prediction method as a function of the number of iterations. The following figure confirms the result shown in Figure 22. After approximately five iterations the error is reduced at a significantly lower rate.

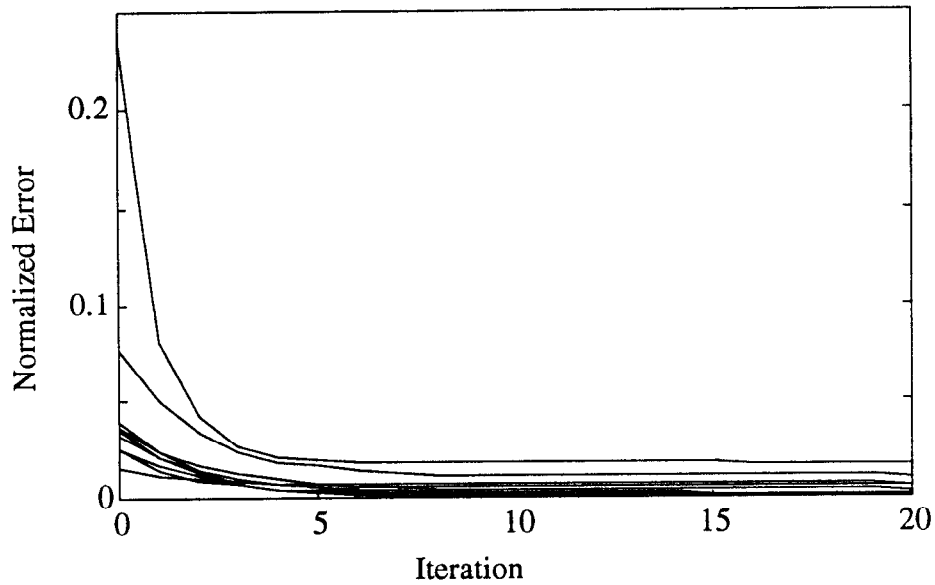


Figure 25. Correlation inversion error for the delta phase method. The error for every tenth channel as a function of the number of iterations is shown. An impulse is used as the input to the cochlear model.

These results show the relative error between the different methods but they do not provide an estimate of the accuracy of the resynthesized signal in any meaningful fashion. In the following section a phrase of speech is resynthesized and the perceptual results are reported.

5 Resynthesizing Speech

As stated in the introduction this work is motivated by the need to resynthesize speech signals from their correlograms. To ascertain if the resynthesis process has adequately reproduced the original signal an informal perceptual test was performed by resynthesizing a short phrase, “A huge tapestry hung in her hallway” (TRAIN/DR5/FCDR1/SX106/SX106.ADC;1) from the TIMIT database (Lamel, Kassel, & Seneff, 1986). To the author the perceptual difference between the original and the resynthesized speech is almost imperceptible. Furthermore, a time-domain comparison between the original and resynthesized waveforms can be made. Figure 26 shows a fraction of the phrase Figure 27 shows a more detailed portion of the same phrase. Here it can be seen that the phrase is well resynthesized. This resynthesis is performed using the delta phase technique. If only the phase technique is used in the resynthesis the quality is degraded. Furthermore, if no phase information is employed then the fidelity of the resynthesized speech is significantly reduced. Although these observations are qualitative in nature, they show that the inversion process is successful and that the attempt to improve the coherency between channels using phase information when performing the correlogram inversion produces a noticeable effect. Furthermore the perceptual results confirm that the error metric used in the previous section provides a reasonable measure of the relative performance of the different techniques.

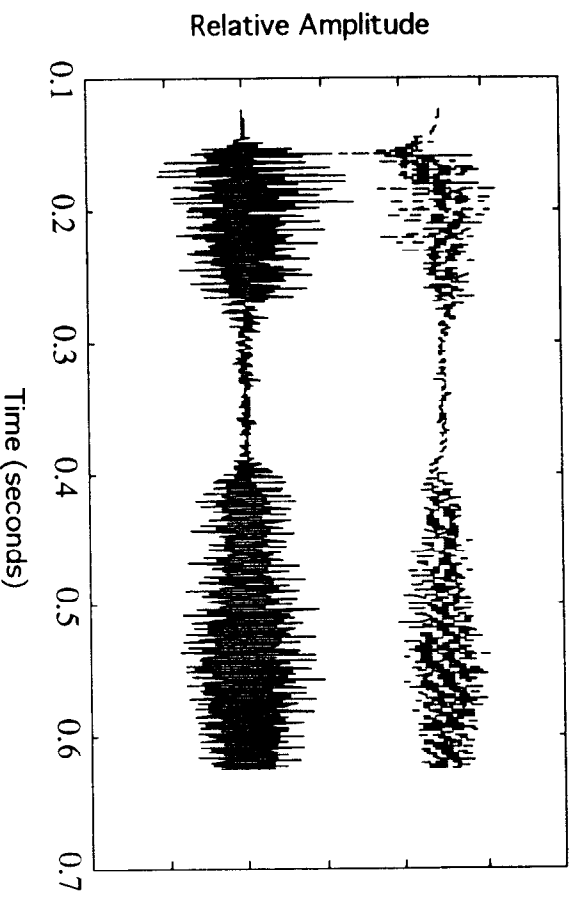


Figure 26. Original signal (bottom) compared to signal resynthesized from its correlogram using the delta-phase technique. Signal is a portion of a sample from the TIMIT database.

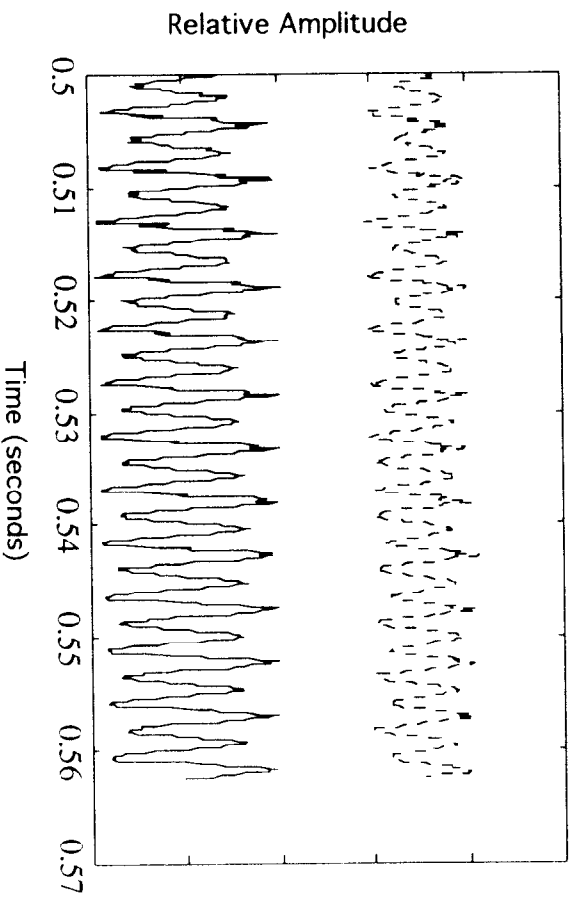


Figure 27. Original signal (bottom) compared to resynthesized signal. Signal is a portion of a sample from the TIMIT database.

6 Conclusion

This thesis explores the problem of resynthesizing a sound from its correlogram. Resynthesizing the original sound requires the inversion of both the cochlear model and the correlogram. The cochlear model inversion results in almost perfect results and no degradation in the sound quality. Inversion from the correlogram itself provides some quite surprising results. Not only is the sound quality barely degraded, but a time-based comparison can be made between the reconstructed signal and the original. These results show that sounds reproduced from the correlogram inversion are excellent reproductions.

In this thesis we have developed a systematic approach to recreate a sound from its correlogram. To reach this goal we have developed a method to resynthesize a sound from its cochleagram and a method to estimate the cochleagram from a correlogram. The method has been tested by resynthesizing a phrase identified as TRAIN/DR5/FCDR1/SX106/SX106.ADC;1 from the TIMIT database. We have shown that the cochleagram inversion process is almost exact. Although an exact inversion of the correlogram has not been achieved, we have been able to invert the correlogram and reproduce the original sound waveform. Informal listening tests by the author show that this inversion has been successful. Thus, the original goals of this thesis have been satisfied.

Although the approach developed in this thesis has proved successful it also requires an inordinate amount of computation. From an engineering perspective, it is also important to examine methods that may not be as exact, but produce adequate results. The method developed here can be easily modified to reduce some of the more intensive computations. For example, the cochleagram inversion can be performed without inverting the half-wave rectification. Understandable results can be achieved without performing a large number of iterations (less than two), however the results will be degraded.

The number of iterations on certain portions of the algorithm can be reduced. However, to test the success it will be necessary to produce a set of object perceptual tests, and that is beyond the scope of this present work. More importantly, however, this approach needs to be applied to partial correlograms. It is anticipated that a method of interpolating between

channels will be necessary to successfully resynthesize a speech signal from a partial correlogram, and again we leave these issues to future work.

REFERENCES

- Duda, R. O., Lyon, R. F., & Slaney, M. (1990). Correlograms and the separation of sound. Proceedings of the 24th Annual Asilomar Conference on Signals, Systems and Computers, 1, 457-461.
- Fisher, W. M., Zue, V., Bernstein, J., Pallett, D. (1987). An Acoustic-Phonetic Data Base. Proceedings of the 113th Meeting of the Acoustical Society of America, May.
- Griffin, D. W., & Lim, J. S. (1984). Signal Estimation from Modified Short-Time Fourier Transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32, 566-578.
- Lamel, L. F., Kassel, R. H., & Seneff, S. (1986). Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. Proceedings of the DARPA Speech Recognition Workshop, Feb., 100-109.
- Licklider, J. C. R. (1951). A Duplex Theory of Pitch Perception. Experientia, 7, 128-133. [Reprinted- Schubert, E.D. (Eds.). (1979). Physiological Acoustics. Stroudsburg, PA: Dowden, Hutchinson and Ross, Inc.].
- Lyon, R. F. (1982). A Computational Model of Filtering, Detection, and Compression in the Cochlea. IEEE Transactions on Acoustics, Speech, and Signal Processing, 12, 1282-1285.
- Oppenheim, A. V., & Schaffer, R. W. (1989). Discrete-Time Signal Processing. Englewood Cliffs, NJ: Prentice Hall.
- Pickles, J. O. (1988). An Introduction to the Physiology of Hearing (2nd ed.). London: Academic Press.
- Roucos, S., & Wilgus, A. M. (1985). High Quality Time-Scale Modification for Speech. Proceedings of the 1985 IEEE Conference on Acoustics, Speech and Signal Processing, 493-496.
- Sachs, M. B., & Young, E. D. (1988). Encoding of Steady-State Vowels in the Auditory

- Nerve: Representation in Terms of Discharge Rate, Journal of the Acoustical Society of America, **66**, 470-479.
- Slaney, M. (1988). Lyon's Cochlear Model (Tech. Rep. No. 25). Apple Computer Inc., Cupertino, California.
- Slaney, M., & Lyon, R. F. (1990). A perceptual pitch detector. Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing, 357-360.
- Slaney, M., & Lyon, R. F. (1993). On the importance of time-A temporal representation of sound, in Visual Representations of Speech Analysis. Eds. Cooke, M., Beet, S., & Crawford, M., John Wiley and Sons, Sussex, England.
- Slaney, M., & Lyon, R. F. (1991). Apple Hearing Demo Reel (Tech. Rep. No. 25). Apple Computer Inc., Cupertino, California.
- Weintraub, M. (1985). A theory and computational model of auditory monaural sound separation. Doctoral dissertation, Stanford University, California.
- Yang, X., Wang, K., & Shamma, S. (1992). "Auditory Representations of Acoustic Signals." IEEE Transactions on Information Theory, **38**, 824-839.