Acoustic signal processing based on the short-time spectrum

by

Michael Wayne Callahan

March 1976

UTEC-CSc-76-209

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC15-73-C-0363.

TABLE OF CONTENTS

ABSTRACT

The frequency domain representation of a time signal afforded by the Fourier transform is a powerful tool in acoustic signal processing. The usefulness of this representation is rooted in the mechanisms of sound production and perception. Many sources of sound exhibit normal modes or natural frequencies of vibration, and can be described concisely in the frequency domain. The human auditory system performs frequency analysis early in the hearing process, so perception is often best described by frequency domain parameters.

This dissertation investigates a new approach to acoustic signal processing based on the short-time Fourier transform, a two dimensional representation which shows the time and frequency structure of sounds. This representation is appropriate for signals such as speech and music, where the natural frequencies of the source change and timing of these changes is important to perception. The principal advantage of this approach is that the signal processing domain is similar to the perceptual domain, so that signal modifications can be related to perceptual criteria.

The mathematical basis for this type of processing is developed, and four examples are described: removal of broad band background noise, isolation of perceptually important speech features, dynamic range compression and expansion, and removal of locally periodic interfering signals.

CHAPTER 1

INTRODUCTION

1.1 Short-Time Frequency Analysis

The representation of a signal by its Fourier transform is central to acoustic signal processing. The usefulness of this transform derives from the mechanisms of sound production and perception. Many sources of sound exhibit normal modes or natural frequencies of vibration, and such phenomena can be described concisely in the frequency domain. There is clear evidence that the human auditory system performs frequency analysis, and perception is often best described by frequency domain parameters.

In signals such as speech and music, the characteristic frequencies of the source change, and perception depends on the timing of these changes. For signals such as these, a modification of the Fourier transform is desired which will show the salient time and frequency structure of the signal. This result can be obtained by introducing a window or weighting function which isolates a short segment of the signal. The window moves along the signal as time progresses, and Fourier analysis is applied to the portion of the signal seen through the window. The result is the short-time Fourier transform [1,2] - a two dimensional, time/frequency representation of the signal:

$$S(\omega,t) = \int_{-\infty}^{\infty} m(t-x)\cdot s(x)\cdot \exp(-i\omega x)\, dx \qquad (1.1)$$

where s(t) is the signal and m(t) is the window. The short-time Fourier transform is invertible [1], and if $m(0) \neq 0$ the inverse is usually chosen as

$$s(t) = [2\pi m(0)]^{-1} \int_{-\infty}^{\infty} S(\omega, t) \cdot \exp(i\omega t)\, d\omega . \qquad (1.2)$$

A visible representation of the short-time Fourier transform of speech is shown in Figures 1 and 2. The ordinate in these figures is frequency (0-5000 Hz), the abcissa is time (1 sec), and the intensity of the reflected light is proportional to the magnitude of $S(\omega, t)$. The difference in appearance of the two figures is due to the window. Figure 1 was made with a relatively short window (10 msec) so that time resolution is good - individual pitch pulses caused by the vocal cords opening and closing can be seen. Figure 2 was made with a longer window (30 msec) and frequency resolution is increased - pitch pulses are not visible but the resulting harmonic structure is. The effect of the window will be discussed in more detail in the next chapter.

Time/frequency displays like Figure 1 and 2, which are commonly called spectrograms, are widely used by speech researchers [2,3]. These displays are useful because they show features important to production and perception of speech. This is illustrated by a series of experiments performed at Haskins Laboratories [4,5]. In these experiments, a mechanical system was used to play back stylized, hand-painted spectrograms of closely related sounds, such as syllables beginning with p, t, and k. It was found that perceptual distinctions among these sounds were explained by differences in their time/frequency structure such as the peak

frequency of the noise burst and the direction of subsequent formant* transitions. These features must be accurately reproduced by systems for synthesis or transmission of speech, but they are not readily apparent in the speech waveform or its conventional Fourier transform.

Because it models speech production and perception, the short-time Fourier transform has been used in an efficient method of speech encoding. The Phase Vocoder [6] performs analysis based on 1.1 (sampled in frequency), and transmits bandlimited signals corresponding to the magnitude and phase of the short-time transform. Speech is reconstructed at the receiver using 1.2. In a more general sense, nearly all channel vocoders are based on short-time frequency analysis [7].

1.2 Time/Frequency Analysis in the Human Auditory System

Helmholtz [8] first proposed a theory of hearing based on frequency analysis of accustic stimuli. This feature of hearing has been observed in numerous psychophysical experiments [9,10], and was put on a firm physiological basis by von Bekesy [11] and Kiang [12].

A schematic diagram of the peripheral auditory system is shown in Figure 3. Sound pressure waves cause vibration of the eardrum or tympannic membrane. These vibrations are transmitted by the ossicular bones of the middle ear, whose main function is impedance transformation between the air medium in the outer ear and the fluid medium of the cochlea. The cochlea is a slender, fluid filled tube,

---

*Formants are the prominant spectral peaks in speech caused by resonances in the vocal tract.

divided into two chambers by the basilar membrane. It is actually coiled like a snail shell (hence its name), but is shown unrolled for clarity in Figure 3. The vibrations transmitted to the cochlear fluid cause motion of the basilar membrane, and this motion is sensed by the auditory nerve.

Von Bekesy [11] showed that acoustic stimuli set up traveling waves on the basilar membrane. The mechanical properties of the membrane are analogous to those of a non-uniform transmission line: waves of a given frequency travel along the membrane until they reach the point resonant at that frequency, and then are rapidly attenuated. Displacement of the membrane is greatest at the point of resonance. As a result, the response of the basilar membrane at a specific point along its length is much like that of a relatively broad bandpass filter. The bandwidth of successive points is roughly constant Q, so frequency resolution is best at low frequencies and time resolution best at high frequencies.

The auditory nerve terminates in hair cells along the length of the basilar membrane. The exact mechanism of the mechanical to neural conversion is not well understood, but Kiang [12] has shown that the firing rate of each neuron is characterized by a response curve similar to that of the basilar membrane. The frequency analysis of the basilar membrane is preserved in the auditory nerve.

The pattern of nerve firings available to higher centers of hearing is thus a two dimensional, time/frequency representation of the acoustic stimulus, similar to the short-time spectrum. The effective bandwidth of the auditory system based on psychophysical data is about 100 Hz for low frequency stimuli (100-350 Hz), and

increases to about 1000 Hz for 5000 Hz stimuli [2,9].

1.3 Contribution of This Research

This paper investigates acoustic signal processing based on the short-time spectrum. The short-time Fourier transform of the signal is calculated as in 1.1, and the magnitude of the transformed signal is modified in an appropriate way. A new signal is then obtained from the modified transform using 1.2.

This approach is suggested by time/frequency analysis in the auditory system, and is appropriate for systems when subjective perception of the output signal is the primary measure of performance. The short-time Fourier transform is used to provide a time/frequency representation because it is computationally effecient. It does not model auditory system processing in detail - for example, frequency analysis in the auditory system is roughly constant Q, while the short-time Fourier transform is constant bandwidth. However, the simplifications inherent in use of this transform do not unduly limit its effectiveness in predicting perception for the experiments described here.

This approach makes use of spectrograms as an aid for system design. The source of perceptual problems can often be recognized by looking at a spectrogram of the signal. For example, one can make a fairly good visual separation of signal and background noise in the spectrogram of a noisy recording shown in Chapter 3. The regions of the spectrogram where noise is visually obvious are generally those most noticeable in listening. An algorithm which selectively attenuates portions of the spectrogram dominated by noise will reduce the noise perceived in the modified signal.

Time/frequency processing results in nonlinear, adaptive systems whose effect is determined by the local time and frequency structure of the signal. These systems are inherently more flexible than linear, time-invariant systems whose design must be based on average or worst case signal characteristics. The disadvantage of these systems is their mathematical complexity. Very few of the systems investigated here can be analyzed in a concise and complete way.

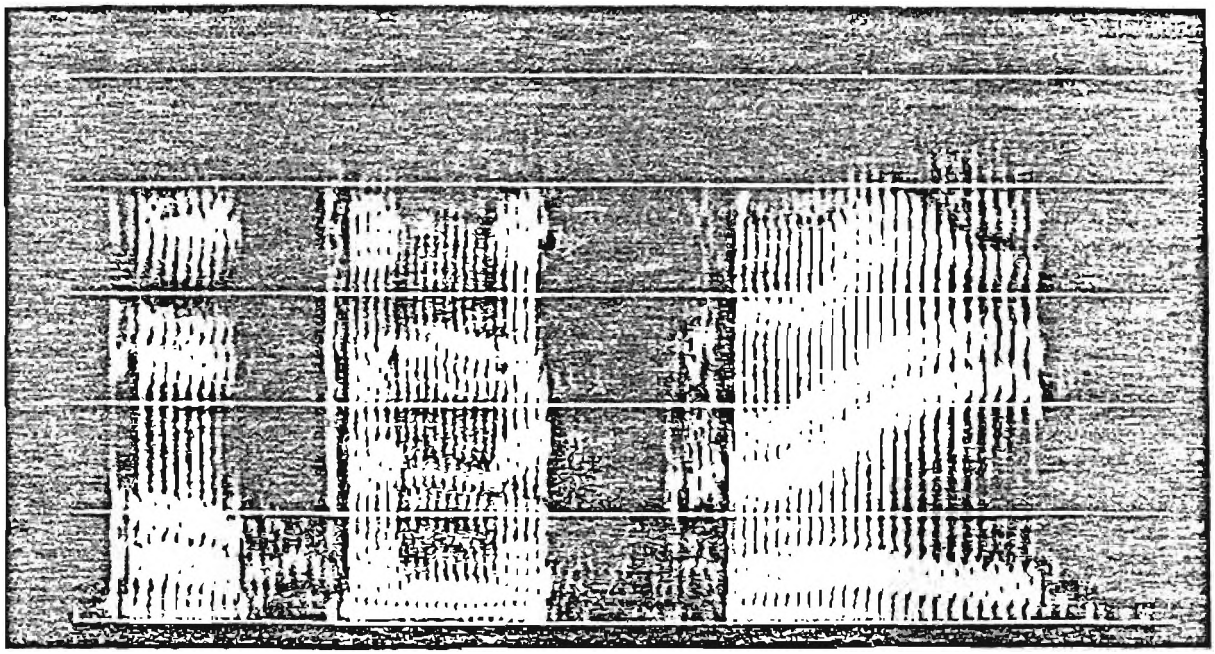


FIGURE 1

Magnitude of the short-time Fourier transform using short (10 msec) window. The picture represents 0.8 sec of speech, 0 - 5000 Hz. It has been scaled 6 dB/oct above 400 Hz to improve the visibility of high frequency components. The speech is "open the crate."

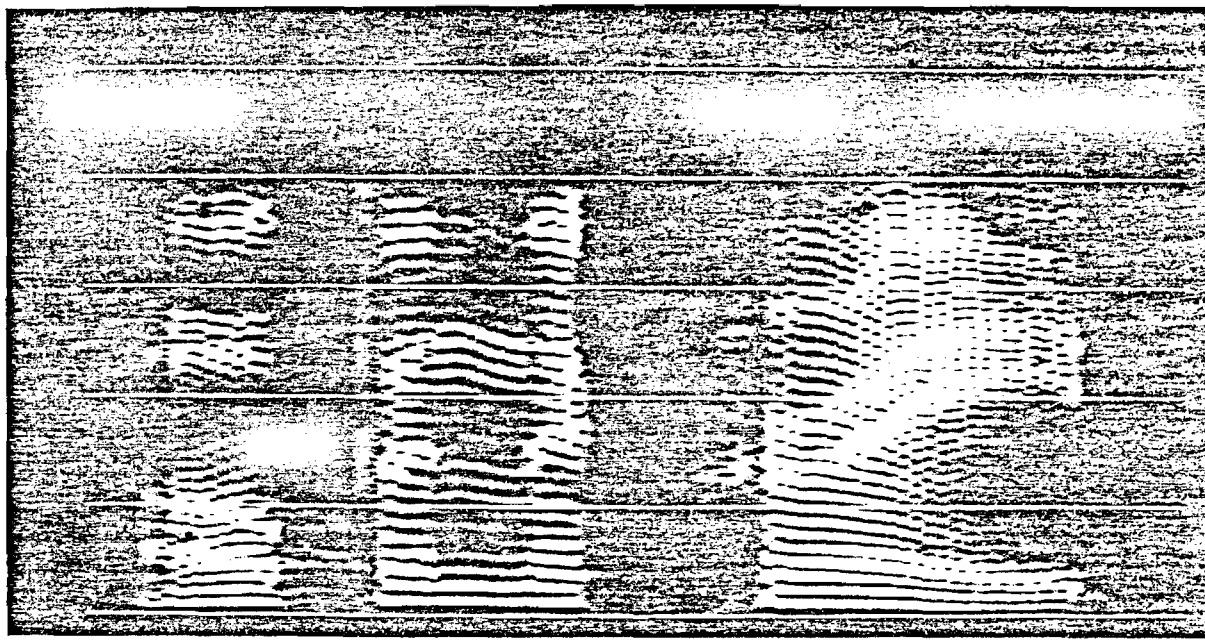


FIGURE 2

Magnitude of the short-time Fourier transform using long (30 msec) window. The speech and other parameters are the same as in Figure 1.

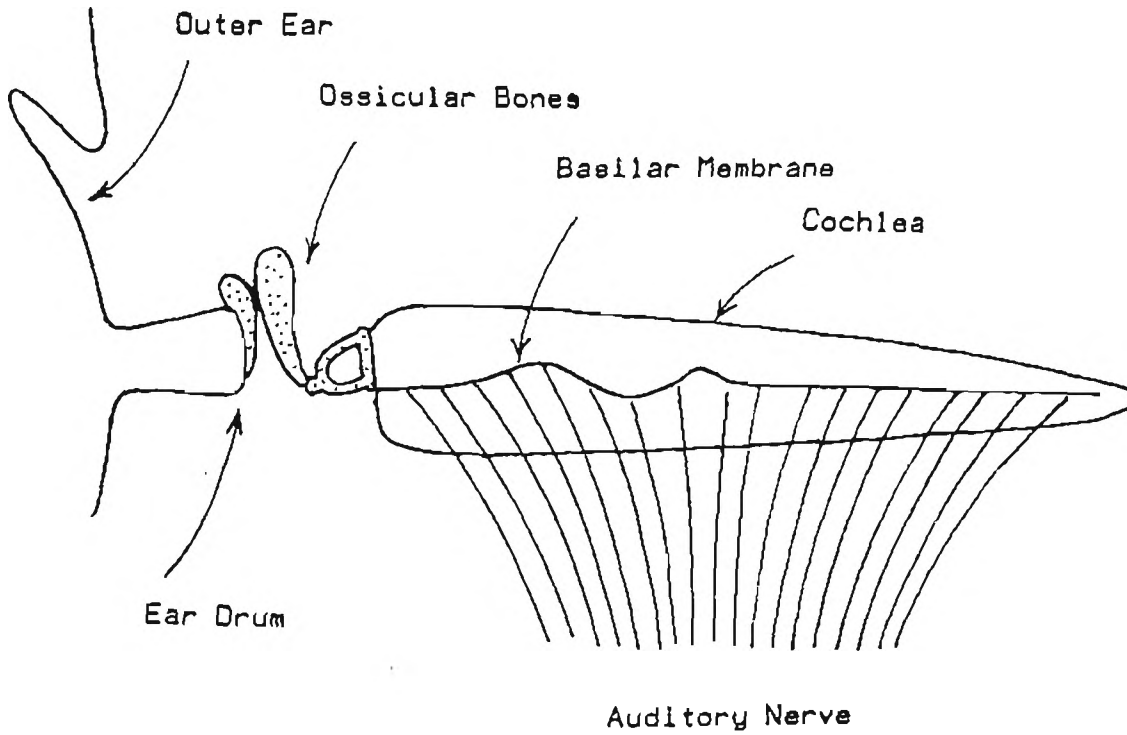


FIGURE 3

Schematic drawing of the peripheral auditory system.

CHAPTER 2

THEORY

2.1 The Fourier Transform of Discrete Signals

The Fourier transform for discrete (sampled) signals is similar to the Fourier series representation of periodic signals, with the role of the time domain and the frequency domain reversed. The time domain signal is discrete, so the frequency domain signal is periodic. The period of the Fourier transform is inversely proportional to the sampling period - this causes the phenomenon of frequency domain aliasing, where successive periods of the Fourier transform overlap if the time signal is not sampled at a sufficiently high rate [13].

If f(n) is a time sequence satisfying

$$\sum_{n=-\infty}^{\infty} |f(n)| < \infty , \tag{2.1}$$

the Fourier transform of f(n) and its inverse are

$$F(\Omega) = \sum_{n=-\infty}^{\infty} f(n)\cdot\exp(-i\Omega_k n), \tag{2.2}$$

$$f(n) = (2\pi)^{-1}\int_{-\pi}^{\pi} F(\Omega)\cdot\exp(i\Omega n)\, d\Omega . \tag{2.3}$$

Ω and n are dimensionless. They are related to the radian frequency and time by the sampling period T:

$$\omega = \Omega/T \quad ; \quad t = nT . \tag{2.4}$$

It can be seen from 2.2 that $F(\Omega)$ has period 2π.

Transforming 2.2 and 2.3 leads to the following two equations, which will be useful later:

$$\delta_{nl} = (2\pi)^{-1}\int_{-\pi}^{\pi} \exp[i\Omega(n-l)]\, d\Omega, \tag{2.5}$$

$$2\pi\cdot\delta(\Omega-\Omega') = \sum_{n=-\infty}^{\infty} \exp[i(\Omega-\Omega')n]. \tag{2.6}$$

δ_{nl} is the Kroniker delta and $\delta(\Omega-\Omega')$ is the (periodic) Dirac delta function.

This brief review is sufficient for the development of the short-time Fourier transform in this chapter. The properties of the Fourier transform of discrete signals are discussed in detail in a number of standard texts, such as References 13 and 14.

2.2 The Short-Time Fourier Transform

The discrete short-time Fourier transform of a sequence $s(n)$ is defined by

$$S(k,n) = \sum_{r=-\infty}^{\infty} s(r)\cdot m(n-r)\cdot\exp(-i\Omega_k r) \tag{2.7}$$

$$\Omega_k = 2\pi k/N \qquad k = 0,1,\cdots N$$

where $m(n)$ is an appropriately chosen window. $S(k,n)$ is a two dimensional, complex function of frequency (k) and time (n). It can be interpreted as samples of the Fourier transform of the portion of the sequence $s(n)$ seen through the sliding window $m(n)$. The window is of finite length, and is normally chosen as the unit sample response of a low pass filter. The discrete short-time spectrum is the squared magnitude of $S(k,n)$.

Since the window $m(n)$ is of finite length, 2.7 can be

implemented efficiently using the Fast Fourier Transform algorithm [15].

The inverse to 2.7 can be chosen as [1]

$$s(n) = 1/N \sum_{k=0}^{N-1} S(k,n) \cdot \exp(i\Omega_k n). \tag{2.8}$$

This inverse imposes a mild restriction on the window, which can be seen by substituting 2.7 into 2.8.

$$s'(n) = 1/N \sum_{k=0}^{N-1} \exp(i\Omega_k n) \cdot \sum_{r=-\infty}^{\infty} s(r) \cdot m(n-r) \cdot \exp(i\Omega_k r) \tag{2.9}$$

Since both sums have a finite number of bounded terms, the order of summation can be reversed.

$$s'(n) = \sum_{r=-\infty}^{\infty} s(r) \cdot m(n-r) \cdot 1/N \sum_{k=0}^{N-1} \exp[i\Omega_k (n-r)] \tag{2.10}$$

$$1/N \sum_{k=0}^{N-1} \exp(i\Omega_k l) = \begin{cases} 1 & l = 0, \pm N, \pm 2N, \cdots \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

Now if the inverse transform is to result in the original signal, the window must satisfy

$$m(n) = \begin{cases} 1 & n = 0 \\ 0 & n = \pm N, \pm 2N, \cdots \end{cases} \tag{2.12}$$

but is otherwise arbitrary.

Notice that with the inverse defined by 2.8, only one point of the time sequence is obtained from each time slice of S(k,n). This is a desireable characteristic for the experiments discussed later, in which S(k,n) is modified and 2.8 is used to obtain a new time signal. Changes to S(k,n) at a given time will be reflected only in the corresponding samples of the time signal.

The short-time transform is normally sampled less densely in

time than the original signal, at a rate determined by the frequency band occupied by the Fourier transform of the window. One might therefore expect that intermediate samples of S(k,n) would have to be obtained by interpolation prior to reconstruction of s(n). For an N point signal and a K point transform desampled in time by a factor L, this would require (N-N/L)·K interpolations* and N calculations using 2.8. Assuming that about 9 multiply and add operations are required for an interpolation, the time required for reconstruction is

$$t_r = [(N-N/L)\cdot 9K + NK]\tau \qquad (2.13)$$

where τ is the multiply and add time. The most distressing feature of 2.13 is the fact that reconstruction time is proportional to the frequency resolution of the short-time transform.

Portnoff [15] has developed a more efficient algorithm to accomplish the interpolation and inverse transform which makes use of the Fast Fourier Trarsform algorithm. Portnoff's method, which is explained in detail in Reference 15, reverses the order in which the interpolation sum and the sum required by 2.8 are performed. This results in a sum which is similar to 2.8, but has the form of a discrete Fourier transform of S(k,n) in the k direction. A new array can therefore be obtained from S(k,n) using the Fast Fourier Transform, and points in the time sequence are obtained by interpolating selected data in the transformed array. Only one

---

*S(k,n) is complex, but due to real-even/imaginary-odd symmetry, only K interpolations are needed for each time slice.

interpolation is required for each intermediate point in the time sequence. The Fast Fourier Transform requires a time proportional to $K \cdot \log_2(K)$* to transform a K point sequence [13]. The constant of proportionality is typically about three times the multiply and add time, so the time required for reconstruction using the Portnoff method is

$$t_2 = [(N/L) \cdot 3K \cdot \log_2(K) + (N-N/L) \cdot 9]\tau . \qquad (2.14)$$

For a given window shape, K/L is a constant - scaling the window length changes the desampling ratio by the same factor. Reconstruction time therefore grows only as the logarithm of the frequency resolution of $S(k,n)$.

To appreciate the time savings afforded by Portnoff's method, consider a case where $K = 512$ and $L = 64$ (typical for some of the experiments described later). Neglecting N/L with respect to N

$$t_1 \simeq 10K \cdot N\tau = 5120 \cdot N\tau ,$$

$$t_2 \simeq (24 \cdot \log_2(K) + 9)N\tau = 225 \cdot N\tau ,$$

$$t_1/t_2 \simeq 23 .$$

2.3 Filter Bank Analogy to the Short-Time Fourier Transform

Equations 2.7 and 2.8 can be interpreted in another way which may provide additional insight concerning the short-time Fourier transform. Consider the system shown in Figure 4. The N bandpass

---

*This assumes that K is a power of 2, which was always the case in the work described here. The Fast Fourier Transform will provide significant time savings whenever K is highly composite [13].

filters BP_0, BP_1, $\cdots$ BP_{N-1}, have unit sample responses

$$b_k(n) = m(n)\cdot\exp(i\Omega_k n) \qquad k=0,1,\cdots N-1 \tag{2.15}$$

where m(n) is the unit sample response of a low pass filter. The filter outputs are therefore

$$y_k(n) = \sum_{r=-\infty}^{\infty} b_k(n-r)\cdot s(r) \tag{2.16}$$

$$= \sum_{r=-\infty}^{\infty} m(n-r)\cdot s(r)\cdot\exp[i\Omega_k(n-r)]. \tag{2.17}$$

The outputs are then multiplied by $\exp(-i\Omega_k n)$, so that all signals will be low pass. This process is often referred to as complex demodulation.

$$z_k(n) = y_k(n)\cdot\exp(-i\Omega_k n) \tag{2.18}$$

$$= \sum_{r=-\infty}^{\infty} m(n-r)\cdot s(r)\cdot\exp(-i\Omega_k r) \tag{2.19}$$

Comparison of 2.19 and 2.7 shows that S(k,n) is equivalent to the demodulated output of a set of bandpass filters. Multiplication by $\exp(-i\Omega_k n)$ does not affect the magnitude, so if we are only interested in |S(k,n)| demodulation is not required.

$$|S(k,n)| = |y_k(n)| = |z_k(n)| \tag{2.20}$$

Reconstruction of a time signal is accomplished by reintroducing the center frequency of each filter and summing. If the output is scaled by 1/N

$$y(n) = 1/N\sum_{k=0}^{N-1} z_k(n)\cdot\exp(i\Omega_k n) \tag{2.21}$$

$$= 1/N\sum_{k=0}^{N-1} y_k(n). \tag{2.22}$$

The system is the identity system and reproduces s(n) if m(n)

satisfies 2.12, as before.

The equivalence of analysis with 2.7 and synthesis with 2.8 to bandpass filtering and summing will be helpful in understanding the experiments discussed later, where the magnitude of S(k,n) is modified prior to reconstruction of a time signal.

2.4 Modification of the Short-Time Fourier Transform

Chapters 3 through 6 discuss experiments which are based on modifying the short-time Fourier transform. Three operations are involved: multiplication of S(k,n) by a function G(k,n) (the local Wiener filter or a threshold function), convolution of |S(k,n)| with a specified function, and convolution of log|S(k,n)| with a specified function. In each case the resulting modified transform is converted to a time signal using 2.8, and this new time signal is reanalyzed - at least by the auditory system. It would be nice to express the effect of each of these operations in a mathematically concise way, and show how the second short-time transform is related to the original. However, only multiplication is mathematically tractible. Fortunately, the results for multiplication provide some insight into the effect of the other operations. In particular, these results show some limitations on the modifications which can be accomplished.

Consider first multiplication of the the short-time transform by a function of frequency only. This represents the limiting case where the spectrum of the input signal changes very slowly. A new time signal is reconstructed with 2.8.

$$s'(n) = 1/N \sum_{k=0}^{N-1} G(k) \cdot S(k,n) \cdot \exp(i\Omega_k n) \qquad (2.23)$$

$$= 1/N \sum_{k=0}^{N-1} \sum_{r=-\infty}^{\infty} G(k) \cdot m(n-r) \cdot s(r) \cdot \exp[i\Omega_k (n-r)] \qquad (2.24)$$

The order of the sums can be reversed, since both are finite and bounded. We define a new function

$$g(l) = 1/N \sum_{k=0}^{N-1} G(k) \cdot \exp(i\Omega_k l) \qquad (2.25)$$

and in terms of g(l), 2.24 becomes

$$s'(n) = \sum_{r=-\infty}^{\infty} g(n-r) \cdot m(n-r) \cdot s(r). \qquad (2.26)$$

Equation 2.26 is a convolution sum, so the original signal has in effect been passed through a filter with unit sample response

$$h(n) = m(n) \cdot g(n). \qquad (2.27)$$

The frequency response of this filter is the Fourier transform of h(n).

$$H(\Omega) = \sum_{n=-\infty}^{\infty} m(n) \cdot g(n) \cdot \exp(-i\Omega n) \qquad (2.28)$$

$$= 1/N \sum_{n=-\infty}^{\infty} \sum_{k=0}^{N-1} m(n) \cdot G(k) \cdot \exp[-i(\Omega - \Omega_k) n] \qquad (2.29)$$

$$= 1/N \sum_{k=0}^{N-1} G(k) \cdot M(\Omega - \Omega_k) \qquad (2.30)$$

Equation 2.30 shows that the intended modification is smoothed by the Fourier transform of the window. The equivalent filter cannot have detail finer than the bandwidth of $M(\Omega)$. This is a manifestation of the time-frequency uncertainty principle discussed in the next section.

Equation 2.30 can be interpreted in terms of the filter bank analogy of Figure 4. G(k) acts like a gain setting for each of the N filters - the frequency response of the k^{th} filter, $M(\Omega - \Omega_k)$, is

multiplied by G(k). The composite frequency response of the filter bank is the sum of the N individual frequency responses.

Equation 2.25 shows that g(n) is periodic, with period N. The unit sample response desired is the principal period of g(n), $0 \le n \le N-1$. When the window is longer than N, 2.27 shows that points from the second period of g(n) are included in h(n). This will produce audible effects in the output signal which are most pronounced when G(k) is not smooth. For this reason, the window length should not exceed the number of frequency samples in S(k,n). This restriction is not limited to multiplication, but applies whenever S(k,n) is modified.

Next we construct the short-time transform of the modified signal. It is convenient to use the filter bank analogy of Figure 4.

$$y_k'(n) = b_k(n) \circledast s'(n) \tag{2.31}$$

$$= b_k(n) \circledast h(n) \circledast s(n) \tag{2.32}$$

$$= y_k(n) \circledast h(n) \tag{2.33}$$

where the symbol $\circledast$ represents convolution. Equation 2.33 can be written in terms of the short-time transform using 2.18.

$$S'(k,n) = \sum_{r=-\infty}^{\infty} S(k,n-r) \cdot h(r) \cdot \exp(-i\Omega_k r) \tag{2.34}$$

The short-time spectrum has been convolved in the time direction with h(n); the term $\exp(-i\Omega_k r)$ compensates for the demodulation of S(k,n). If G(k) and thus H(Ω) are approximately constant over frequency ranges comparable to the bandwidth of H(Ω), synthesis and reanalysis will be approximately an identity operation.

$$S'(k,n) \simeq G(k)\cdot S(k,n) \qquad (2.35)$$

Next, consider multiplication of S(k,n) by a function of time and frequency. In this case, 2.31 becomes a superposition sum

$$s'(n) = \sum_{r=-\infty}^{\infty} g(n-r,n)\cdot m(n-r)\cdot s(r) \qquad (2.36)$$

where

$$g(l,n) = 1/N \sum_{k=0}^{N-1} G(k,n)\cdot \exp(i\Omega_k l). \qquad (2.37)$$

Each sample of s'(n) is now computed from a new unit sample response. The intended modification G(k,n) is smoothed in frequency by M(Ω), as before.

Defining $h(l,n) = g(l,n)\cdot m(n)$, the short-time transform of the modified signal is

$$S'(k,n) = \sum_{r=-\infty}^{\infty}\sum_{x=-\infty}^{\infty} m(n-r)\cdot h(r-x,r)\cdot s(x)\cdot \exp(-i\Omega_k r). \qquad (2.38)$$

We cannot proceed as in 2.31 - 2.34 because of the time dependence of h(l,n). The physical reason can be seen by considering the filter bank analogy. Multiplication by G(k,n) is equivalent to a time-varying gain for each filter. The output of each filter is amplitude modulated, which broadens the signal spectrum so that it will not "fit" in the corresponding filter when analyzed the second time. However if h(l,n), viewed as a function of n, is approximately constant over the length of the window (i.e., the unit sample response of the equivalent filter changes slowly)

$$m(n-r)\cdot h(r-x,r) \simeq m(n-r)\cdot h(n-r,n).$$

Equation 2.38 then becomes

$$S'(k,n) \simeq \sum_{r=-\infty}^{\infty}\sum_{x=-\infty}^{\infty} m(n-r)\cdot h(r-x,n)\cdot s(x)\cdot\exp(-i\Omega_k r). \quad (2.39)$$

Substituting $l = r-x$ and using 2.7

$$S'(k,n) \simeq \sum_{l=-\infty}^{\infty} S(k,n-l)\cdot h(l,n)\cdot\exp(-i\Omega_k l). \quad (2.40)$$

Equation 2.40 is the time dependent result analagous to 2.34.

In order for analysis and synthesis to be an approximate identity, modifications to the short-time transform must change slowly in time, compared to the window, and slowly in frequency, compared to the Fourier transform of the window. This restriction applies whether the modifications are produced by multiplication or some other operation on the short-time transform.

In the case where the modified signal is reanalyzed the auditory system, the interpretation of 2.40 is slightly different. $S(k,n)$ and $S'(k,n)$ represent signals in the auditory system, and not the short-time transform used to produce $h(l,n)$. Assuming that $h(l,n)$ represents the desired modification with sufficient accuracy, it is only necessary that $h(l,n)$ change slowly compared to the effective auditory system window for the modification to be perceived correctly. The auditory system is relatively broad band compared to the short-time transforms used for the experiments described later, so its effective window is shorter [2].

2.5 Selection of a Window

The short-time Fourier transform is really a family of transforms, corresponding to all the possible window functions which can be used in 2.7. Members of this family can be quite different, as is shown by Figure 5. The same input speech was used for all

three spectrograms. Figure 5(a) was made with a Fourier window of length (2NT) 25.6 msec.

$$m(n) = 1 \quad -N < n < N \qquad = 0 \quad \text{otherwise} \tag{2.41}$$

Figure 5(b) was made with a raised cosine or hanning window of the same length.

$$m(n) = .5 + .5\cdot\cos(\pi n/N) \quad -N < n < N \qquad = 0 \quad \text{otherwise} \tag{2.42}$$

Figure 5(c) was made with a hanning window 12.8 msec long. The differences in the appearance of the three spectrograms are due to the shape and length of the window functions and their Fourier transforms. It is obvious from Figure 5 that that the effect of a process based on the short-time spectrum will depend on the window.

The short-time Fourier transform cannot have arbitrarily good time and frequency resolution. The time and frequency resolution are determined by the window (see Appendix A) and are limited by the well known Heisenberg uncertainty principle [16], which relates the temporal extent of a function $f(t)$ and the frequency spread of its Fourier transform $F(\omega)$.

$$\Delta^2 t = \frac{\int_{-\infty}^{\infty} (t-t_0)^2\cdot|f(t)|^2\,dt}{\int_{-\infty}^{\infty} |f(t)|^2\,dt} \tag{2.43}$$

$$\Delta^2 \omega = \frac{\int_{-\infty}^{\infty} (\omega-\omega_0)^2\cdot|F(\omega)|^2\,d\omega}{\int_{-\infty}^{\infty} |F(\omega)|^2\,d\omega} \tag{2.44}$$

$$\Delta t\,\Delta\omega \geq .5 \tag{2.45}$$

When the desired process must change rapidly in both frequency and time, probably the best solution is to match the frequency and

time resolution of the window to the average resolution of the auditory system. In the experiments described here, the process changes slowly compared to the temporal resolution of the ear. In this case, the length of the window is based on the time that the signal is approximately stationary, in order to best use the time/frequency resolution available.

If the window is much longer than the stationary time of the signal, the temporal extent of events is not well resolved and the short-time spectrum begins to represent an average of sequential events. If the window is much shorter, the frequency resolution of the short-time spectrum is unnecessarily reduced. For signals such as speech and music, the stationary time varies and the shape of the window must be chosen to represent spectral features of interest well over a range of stationary times. The usual assumption for speech is that the signal is stationary for about 20 msec. This corresponds approximately to the length of the window used for Figures 5(a) and 5(b).

The minimum uncertainty function which satisfies the equality in 2.45 is the gaussian. Landau and Pollock [17] have solved the uncertainty problem under conditions more appropriate to the discrete short-time Fourier transform. They show that the function which vanishes for $|t| > T/2$ and has as much of its energy as possible in the interval $|\omega| < \Omega$ is the zeroth prolate spheroidal wave function, $\psi_0(t, \Omega T/2)$. Many other finite length window functions for spectral analysis have been investigated [18,19]. Each solves the problems of spectral resolution and computational efficiency in a slightly different way. Most of these functions are

smooth.

A hanning window was used for all the experiments described in this paper. This window was easy to implement and provided adequate time and frequency resolution. It is likely that some improvement could be made in each experiment by optimizing the window shape for the needs of the particular process. However, informal tests using other window shapes indicate that, so long as the window is smooth, the experimental results will not change significantly.

Equation 2.12 places no restrictions on window length, so long as the window has appropriate zeros. For processes which modify the short-time spectrum, the window length cannot be greater than the number of frequency samples in S(k,n), as was shown in the last section.

2.6 Information in the Magnitude and Phase of the Short-time Fourier Transform

The remaining chapters describe processes which alter the magnitude of S(k,n) but leave the phase unchanged. An experiment was performed to determine how information in speech is divided between the magnitude and phase, and indicate limitations inherent in processing the magnitude only.

The short-time transform of a typical sample of speech was calculated using a 25.6 msec hanning window. Two new signals were then synthesized from this data using 2.8. The first used the original phase and flat magnitude. The second used the original magnitude and a random phase function. The resulting signals are shown in Figure 6. The magnitude of their short-time transforms are

shown in Figure 7.

The flat magnitude, original phase signal contains mostly excitation information. One can tell whether the speech is voiced or unvoiced and determine the pitch, but the speech is otherwise unintelligible. Pitch pulses can be seen in the waveform of Figure 6(b), and harmonic structure is evident in the spectrogram of Figure 7(b). The inability of magnitude flattening to eliminate pitch structure is to be expected because of the frequency smoothing discussed in the last section. The intended modification, $[|S(k,n)|]^{-1}$, is smoothed by the Fourier transform of the window as in 2.30, and fine detail - primarily harmonic structure due to pitch - is averaged out. It is apparent from Figure 7(b) that smoothing has also occurred in the time direction. This is due to the finite time resolution of the window used to calculate the original short-time transform.

The original magnitude, random phase signal contains mostly vocal tract information. The speech is quite intelligible, but sounds whispered. This is evident from the noise-like structure of the waveform in Figure 6(c), and the lack of harmonic structure in Figure 7(c).

These results are consistent with comments by Flanagan [6], who notes that in a narrow band implementation of the Phase Vocoder the phase signal is primarily excitation and the magnitude signal is primarily vocal tract information.

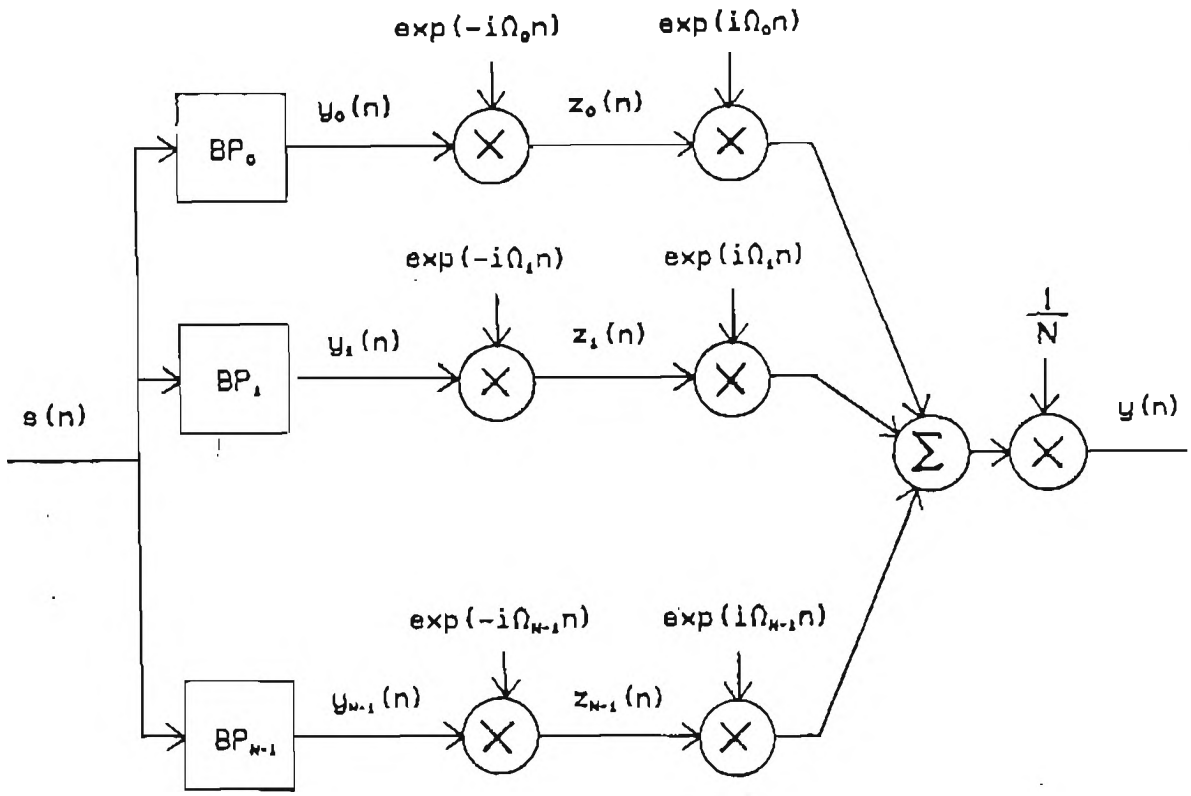
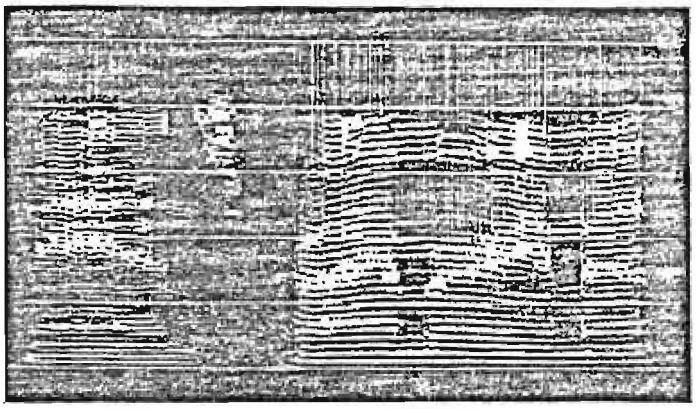


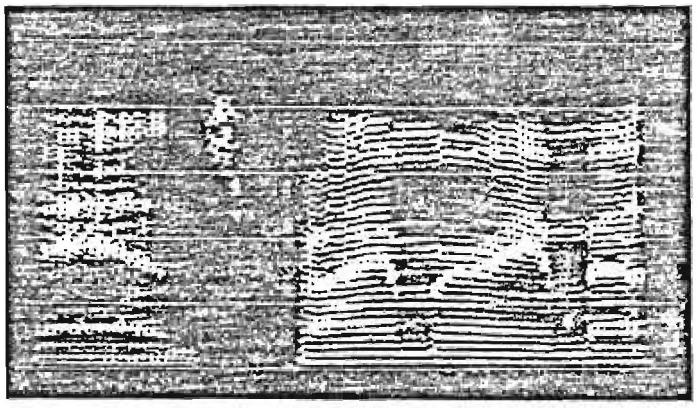
FIGURE 4

Filter bank analogy to the short-time Fourier transform. The signals $z_k(n)$ are equivalent to $S(k,n)$.

(a)



(b)



(c)

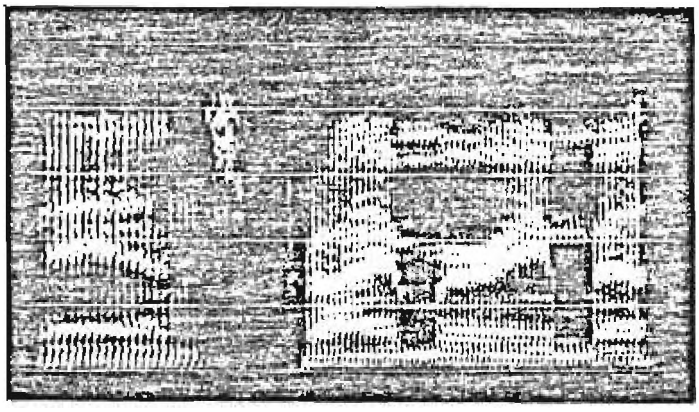
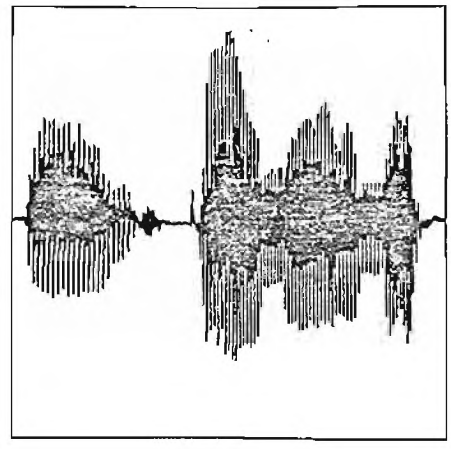
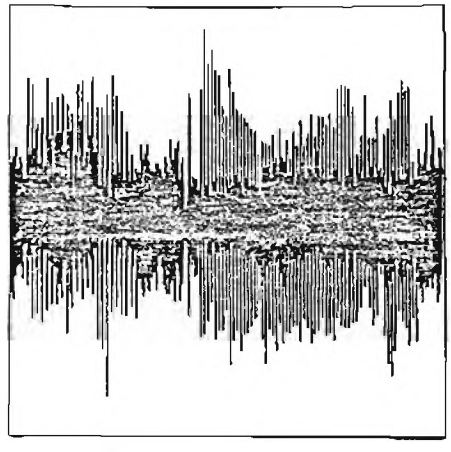


FIGURE 5

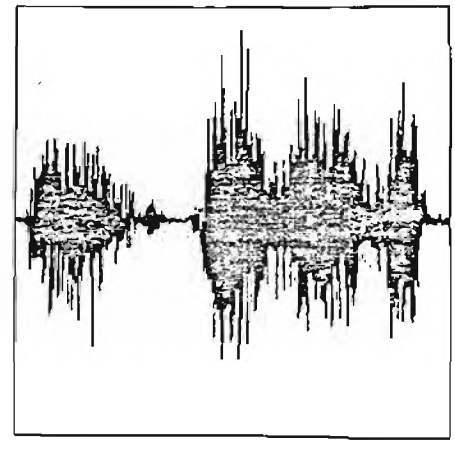
Effect of the window on the short-time Fourier transform: (a) 25.6 msec Fourier window, (b) 25.6 msec hanning window, (c) 12.8 msec hanning window. Each spectrogram represents one second of speech, 0 - 5,000 Hz. All three have been scaled by 6 dB/oct above 400 Hz. The speech is "man's primary."



(a)



(b)

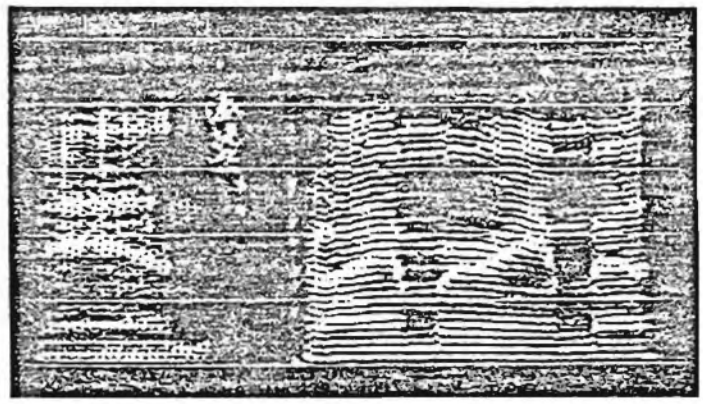


(c)

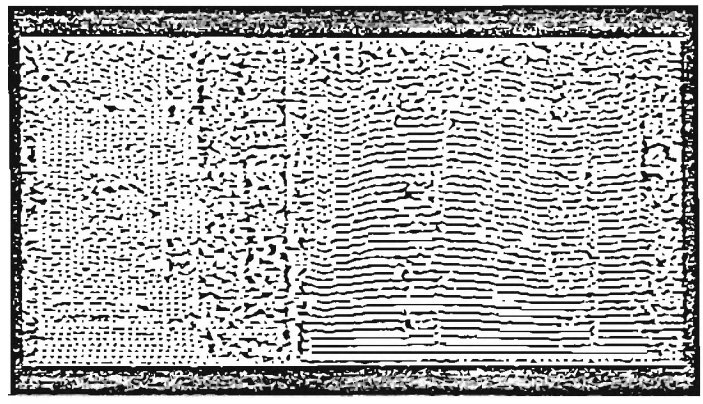
FIGURE 6

Signals reconstructed from the magnitude and phase of the short-time Fourier transform: (a) original speech, (b) flat magnitude, original phase reconstruction, (c) original magnitude, random phase reconstruction. The waveforms are one second long.

(a)



(b)



(c)

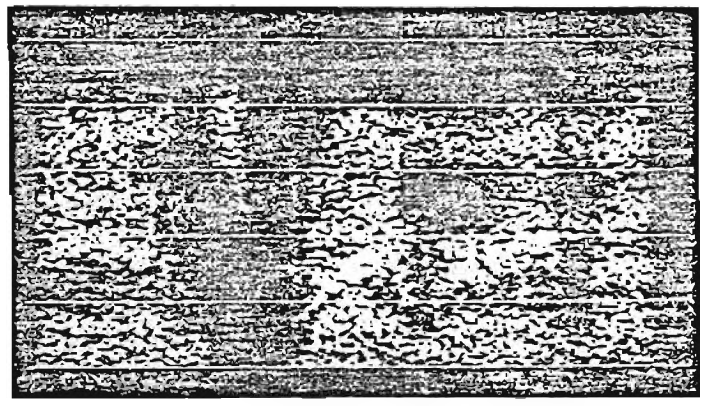


FIGURE 7

Spectrograms of signals reconstructed from the magnitude and phase of the short-time Fourier transform: (a) original speech, (b) flat magnitude, original phase reconstruction, (c) original magnitude, random phase reconstruction. The spectrograms in (a) and (c) have been scaled 6 dB/oct above 400 Hz.

CHAPTER 3

REMOVAL OF BROAD BAND BACKGROUND NOISE

3.1 Local Wiener Filtering

The separation of a signal from additive random noise is a common problem in acoustic signal processing. The optimum linear system for performing such a separation, based on a mean square error criterion, was obtained by Wiener [20]. Given the signal plus noise, the signal is estimated with a linear combination of the data*.

$$x(t) = s(t) + n(t) \tag{3.1}$$

$$\text{est}\{s(t)\} = \sum_{r=-\infty}^{\infty} h(t,r)\cdot x(r) \tag{3.2}$$

If the signal and the noise are stationary, $h(t,r)$ depends only on the time difference $\tau = t-r$.

$$\text{est}\{s(t)\} = \sum_{r=-\infty}^{\infty} h(t-r)\cdot x(r) \tag{3.3}$$

The function $h(\tau)$ is chosen to minimize the mean square error

$$E\{|s(t) - \sum_{r=-\infty}^{\infty} x(t-r)\cdot h(r)|^2\} = \text{minimum} \tag{3.4}$$

where E is the expected value operator. Setting the variation of 3.4 to zero

---

*The discrete time index t is used in lieu of n in this chapter to avoid confusion with the noise signal.

$$\delta[E\{|s(t) - \sum_{r=-\infty}^{\infty} x(t-r)\cdot h(r)|^2\}] = 0 \qquad (3.5)$$

$$E\{[s(t) - \sum_{r=-\infty}^{\infty} x(t-r)\cdot h(r)]\cdot x(t-l)\}\cdot \delta h(l) = 0 \qquad (3.6)$$

we get, since the variation is arbitrary

$$R_{sx}(l) - \sum_{r=-\infty}^{\infty} R_{xx}(l-r)\cdot h(r) = 0 . \qquad (3.7)$$

$R_{sx}(l)$ is the cross correlation $E\{s(t)\cdot x(t-l)\}$, and $R_{xx}(l)$ is the autocorrelation $E\{x(t)\cdot x(t-l)\}$. Taking the Fourier transform of 3.7

$$H(\omega) = \Phi_{sx}(\omega)/\Phi_{xx}(\omega). \qquad (3.8)$$

$\Phi_{sx}(\omega)$ is the cross spectrum of $s(t)$ and $x(t)$, defined as the Fourier transform of R_{sx}, and $\Phi_{xx}(\omega)$ is the spectrum of $x(t)$. If the signal and noise are uncorrelated, $R_{sx} = R_{ss}$ and $R_{xx} = R_{ss} + R_{nn}$, so 3.8 becomes

$$H(\omega) = \frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \Phi_{nn}(\omega)} . \qquad (3.9)$$

Equation 2.9 assumes that the signal is stationary, and does not apply to signals such as speech and music. However, we might envision a running determination of 2.9, based on a local section of the signal which is approximately stationary.

$$H(\omega, t) = \frac{\Phi_{ss}(\omega, t)}{\Phi_{ss}(\omega, t) + \Phi_{nn}(\omega)} \qquad (3.10)$$

A filtering process based on 3.10 can be implemented conveniently using the short-time Fourier transform. The local spectrum of the signal plus noise can be estimated by time averaging the squared magnitude of the short-time transform

$$\Phi_{xx}(k,t) \simeq (2L+1)^{-1} \sum_{l=-L}^{L} |X(k,t+l)|^2 \qquad (3.11)$$

where L is determined by the approximate stationary time of the signal. The noise spectrum can be estimated in the same manner during a period when there is no signal. The noise spectrum estimate will normally have smaller variance than the local spectrum estimate because the noise is stationary and the estimating time can be quite large. Using these estimates and assuming that the signal and the noise are uncorrelated, a local Wiener filter is obtained.

$$H(k,t) = \frac{\Phi_{xx}(k,t) - \Phi_{NN}(k)}{\Phi_{xx}(k,t)} \qquad \Phi_{xx} > \Phi_{NN} \qquad (3.12a)$$

$$H(k,t) = 0 \qquad \Phi_{xx} \leq \Phi_{NN} \qquad (3.12b)$$

The restriction to positive values is required because Φ_{xx} and Φ_{NN} are estimated spectra with finite variance.

This process was used to reduce the surface noise on a 1907 recording by Enrico Caruso. The spectrum of the surface noise was estimated from silent grooves at the beginning and end of the record. Observation of portions of the short-time spectrum which are dominated by noise showed that the noise is stationary throughout the recording. Figure 7 shows the estimated noise spectrum. The average spectrum of the entire recording is also shown for comparison.

The local spectrum of the noisy signal was estimated as in 3.11. A 25.6 msec hanning window was used for the short-time Fourier transform. The averaging time (2L+1) was 100 msec. This time is somewhat longer than the stationary time suggested by the singing, but was the minimum necessary to obtain a local spectrum

estimate with acceptable variance. The short-time transform was multiplied by H(k,t) from 3.12, and a new signal constructed from 2.8.

$$X'(k,t) = H(k,t) \cdot X(k,t) \tag{3.13}$$

$$x'(t) = 1/N \sum_{k=0}^{N-1} X'(k,t) \cdot \exp(i\Omega_k t) \tag{3.14}$$

Spectrograms of a portion of the original and processed signal are shown in Figure 8. The rapidly varying component of the spectrograms is the singing; the relatively constant harmonic structure below ~1500 Hz is the orchestra.

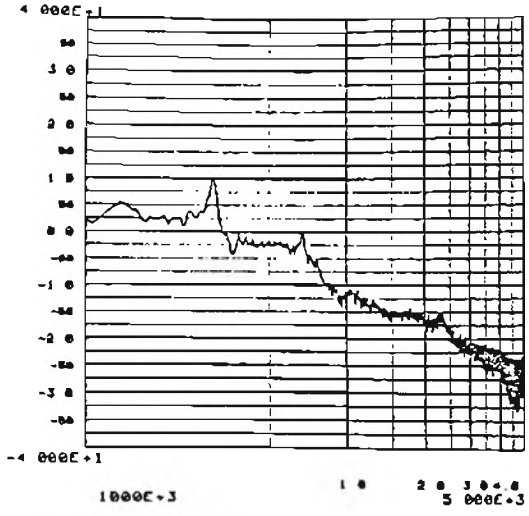
As is apparent in Figure 8(b), the process greatly reduces the surface noise. Careful listening shows no change in the singing voice. The variance of the local spectrum estimate, however, allows narrow bands of noise to be audible, even though they have been greatly attenuated. This effect is quite noticeable in passages where the signal is quiet or relatively narrow band, so that the remaining background noise is not masked. It is more noticeable than white noise at the same level because the noise spectrum is narrow and changes quite rapidly.

3.2 Thresholding

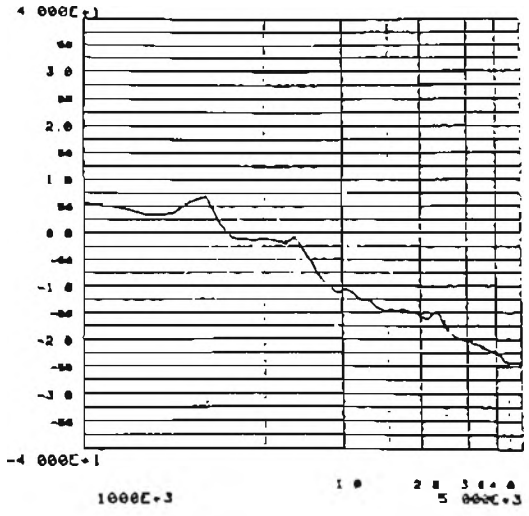
Figure 7 shows that the overall signal to noise ratio for the original recording is about 24 dB. The local signal to noise ratio in $|X(k,t)|^2$ is often better, because the signal is concentrated in frequency and time, while the noise is relatively smooth. The regions of $|X(k,t)|^2$ can therefore be separated from those dominated by noise by a thresholding operation.

The thresholding algorithm was designed to ensure that background noise was attenuated uniformly and to avoid artifacts in the singing such as sudden starts and stops. Two reference levels were used - a lower level 3 dB above the noise spectrum and an upper level 12 dB above noise. When $|X(k,t)|^2$ exceeded the upper threshold, it was left unattenuated (both in the time direction and the frequency direction) until it passed below the lower threshold; otherwise, $|X(k,t)|^2$ was attenuated by 24 dB. The effect can be visualized in terms of a contour map of $|X(k,t)|^2$. If a peak in $|X(k,t)|^2$ is higher than the contour corresponding to the upper threshold, all portions of the peak down to the contour corresponding to the lower threshold are left unchanged. Other regions of the contour map are attenuated. This description must be modified in one respect: to keep computer memory requirements small, the algorithm could look backward in time only 50 msec.

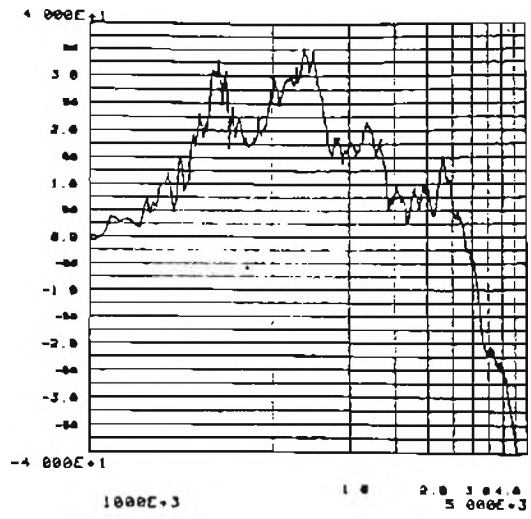
A spectrogram of the resulting signal is shown in Figure 8(c). The thresholding algorithm performed better than the local Wiener filter in that background noise was completely eliminated. Again there was no noticeable change in the singing voice. The only apparent degradation was in the sound of the accompanying orchestra, which was somewhat thin in places because orchestral harmonics remain too near the level of the noise, and are attenuated.



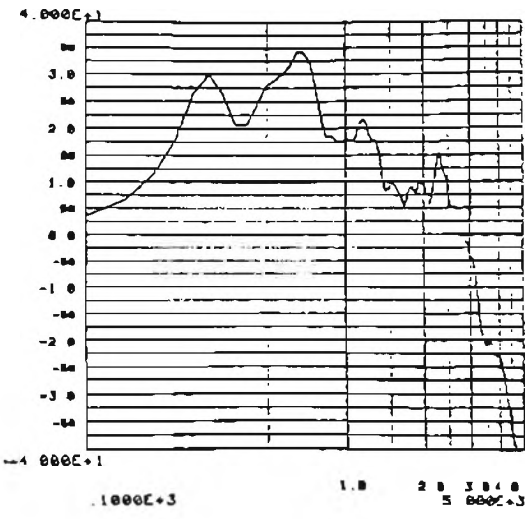
(a)



(b)



(c)

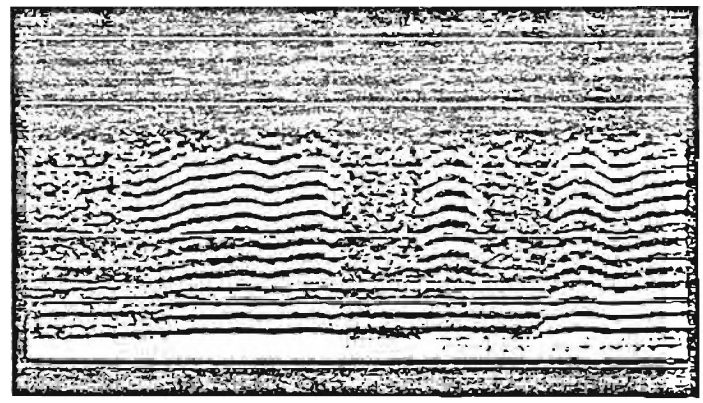


(d)

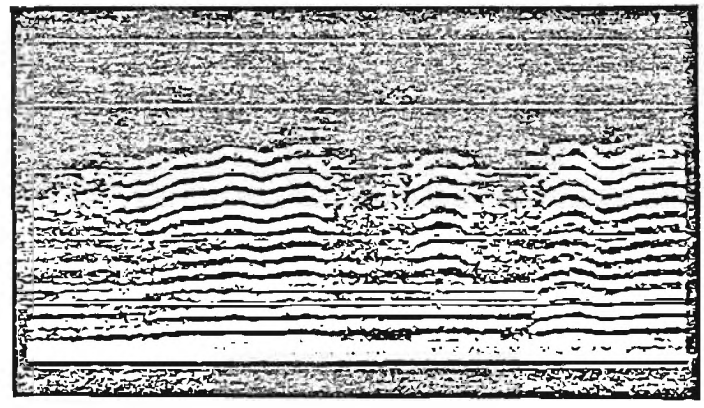
FIGURE 8

Spectrum estimate for Caruso recording: (a) 1024 point noise spectrum estimate, (b) 128 point estimate actually used for processing, (c) 1024 point estimate of average spectrum of entire recording, (d) 128 point average spectrum estimate.

(a)



(b)



(c)

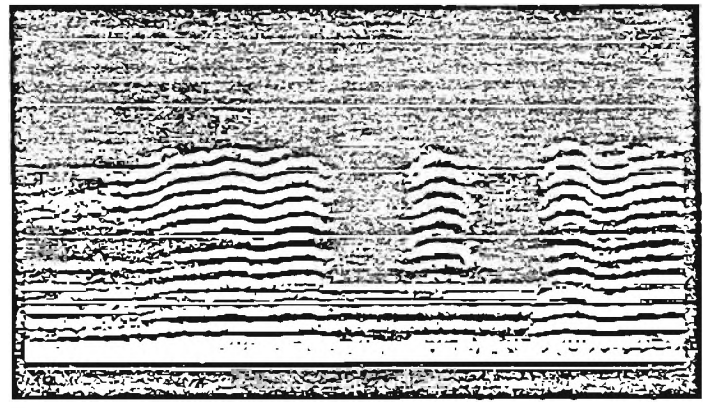


FIGURE 9

Background noise removal: (a) original recording, (b) processed by local Wiener filtering, (c) processed by thresholding. All three spectrograms have been scaled 6 dB/oct above 400 Hz.

CHAPTER 4

ISOLATION OF PERCEPTUALLY IMPORTANT SPEECH FEATURES

There is considerable evidence that certain features of the short-time spectrum are important to perception. Flanagan [2] describes examples of the importance of pitch changes, formant frequencies and bandwidths, and voiced-unvoiced changes. All of these are evident in the short-time spectrum. Liberman's work [5] in synthesizing speech from spectrograms is also an illustrative example: he shows, for example, that in plosive consonants such as p, t, and k, the direction of the formant transition following the noise burst often determines which consonant is perceived.

Experiments were conducted to attempt to isolate selected speech features in $\log|S(k,n)|$, and determine which regions were most important to perception. Three features were selected as typical - pitch, formants, and plosive noise bursts. The logarithm was taken to model the approximate logarithmic sensitivity of the auditory system. In addition, it was desired to isolate these features by two dimensional filtering. Bandpass filtering of the magnitude or squared magnitude of $S(k,n)$ produces functions which are not everywhere positive, and this limits its application to speech processing.

A 25.6 msec hanning window was used for calculation of the short-time Fourier transform. The window was augmented with zeros

to twice this length to increase frequency direction sampling of $S(k,n)$ and minimize aliasing in $\log|S(k,n)|$. The normal time sampling of $S(k,n)$ for this window (sampling period 3.2 msec) was sufficiently conservative to limit time direction aliasing of $\log|S(k,n)|$. The two dimensional filters used were zero phase with flat passbands. Half cosine transition regions were used to minimize ringing effects.

The results of these experiments are summarized in Figure 10, which shows the approximate regions of the Fourier transform of $\log|S(k,n)|$ occupied by pitch information, formant information, and plosive information. The distinction between steady state formants and formant transitions in Figure 10 is somewhat arbitrary. This reflects the fact that there are few sections of speech which are clearly steady state. One should keep in mind that only the low frequency component of the speech features represented in Figure 10 is accessible to processing based on the magnitude only, because of the smoothing effect discussed in Chapter 2.

The regions shown ir Figure 10 generally represent male speech. For female speech, the pitch period sometimes decreases below 3.0 msec so the lower boundary for pitch features extends into the formant region.

Figures 11 and 12 show speech features obtained by two dimensional filtering of $\log|S(k,n)|$. These features were obtained by multiplying the Fourier transform of $\log|S(k,n)|$ by a filter corresponding to the appropriate region in Figure 10, inverse transforming, and exponentiating. Each of the pictures has been scaled to use the full dynamic range of the film, so the figures do

not indicate the amplitude of the features relative to the original speech. The pitch, formants, and plosives have much lower dynamic range than the original speech. Most of the dynamic range of the original is in the low frequency area of Figure 10. This is shown in Figure 12(c), which shows the features obtained by high pass filtering with zero low frequency gain and transitions at $f_k = .5$ msec, $f_n = 25$ Hz.

An interesting feature of these results is that the slowly changing component of $\log|S(k,n)|$ seems to be of lesser importance to perception. This is supported by characteristics of the auditory system. The signal in the auditory nerve exhibits "temporal adaptation" [10,12] - that is, the auditory system adapts quite rapidly (a few tens of milliseconds) to changes in stimulus intensity. This removes the slowly changing temporal component and reduces the dynamic range of the neural signal. The auditory system also exhibits "two-tone inhibition" [9,21] - the threshold for hearing a test tone is increased by nearby tones. This may attenuate the component of the signal in the auditory system which changes slowly with frequency, similar to lateral inhibition in the visual system [22]. It is interesting to note that effects which attenuate the slowly changing component in a signal can be modeled mathematically by high pass filtering the logarithm of the signal magnitude [23], as was done in Figure 12(c).

This discussion suggests that a signal synthesized from Figure 12(c) should have low dynamic range but still be highly intelligible, and this is in fact the case. Speech processed to attenuate the low frequency component of the short-time spectrum

might therefore be more intelligible than normal speech in a noisy environment. Informal listening tests and the results of the next chapter support this notion.

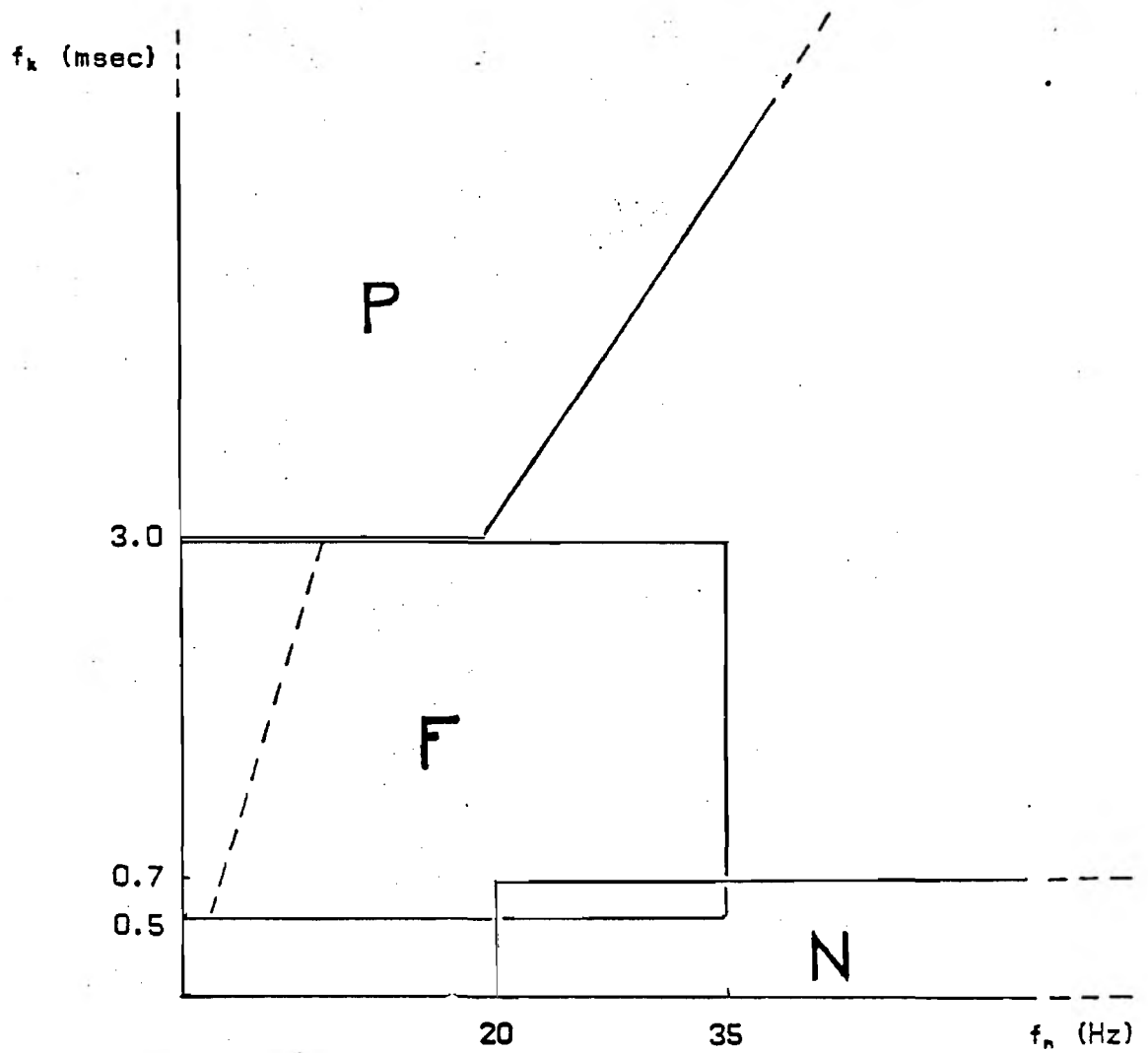
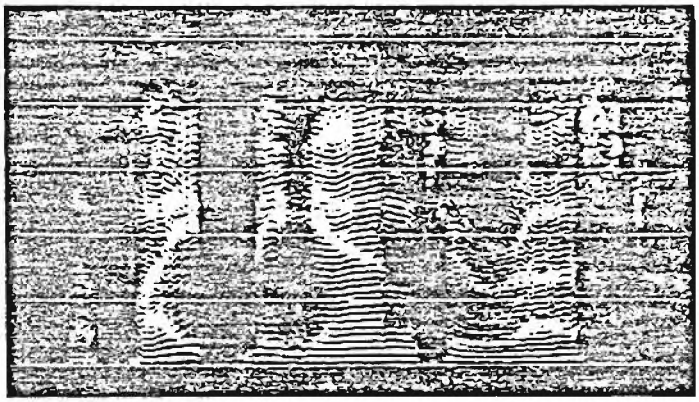


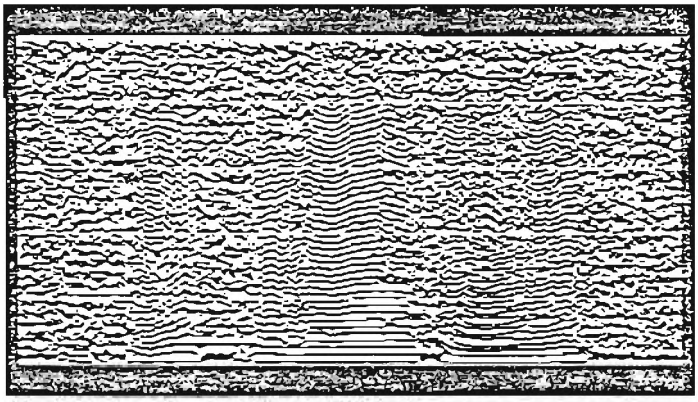
FIGURE 10

Areas of the Fourier transform of $\log|S(k,n)|$ occupied by speech features: F represents formants, P pitch, and N plosive noise bursts. The dotted line represents an approximate division between steady state formants and formant transitions.

(a)



(b)



(c)

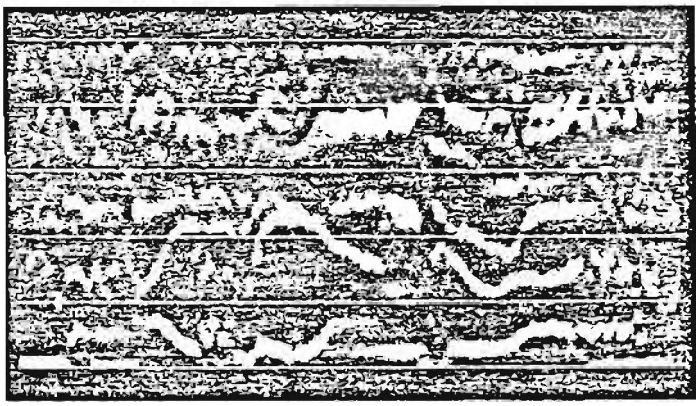
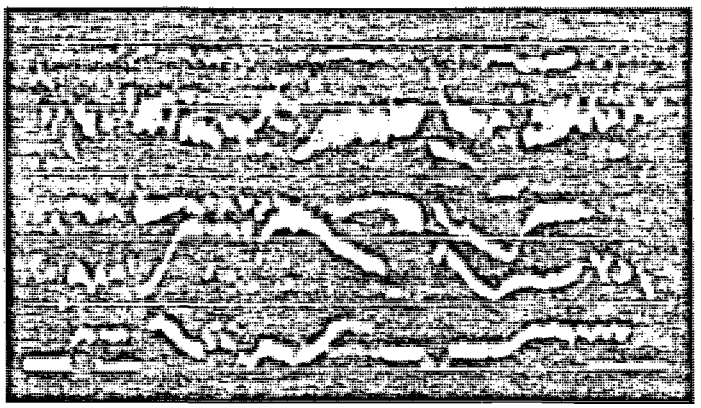


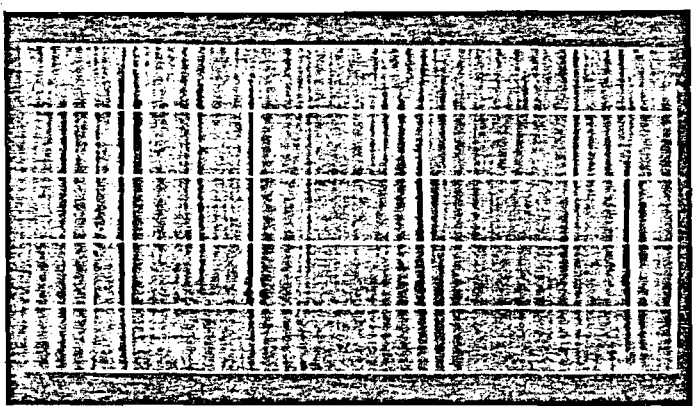
FIGURE 11

Speech features obtained by two dimensional filtering of $\log|S(k,n)|$: (a) original speech, (b) pitch, (c) formants. Each picture represents 1.6 seconds of speech, and has been scaled to use the entire dynamic range of the film. The speech is "the pipe began to rust." The spectrogram in (a) has been scaled 6 dB/oct above 400 Hz.

(a)



(b)



(c)

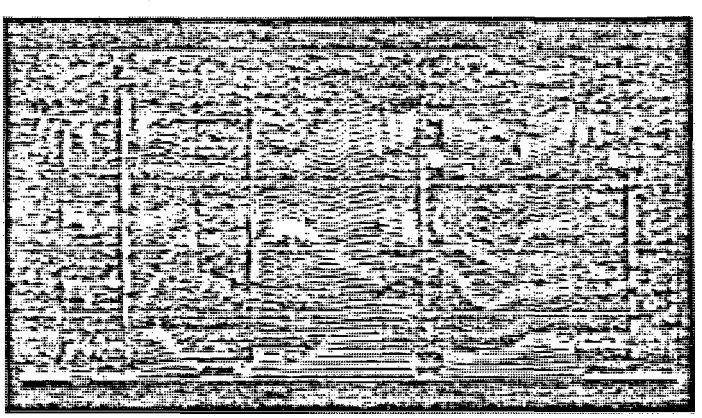


FIGURE 12

Speech features obtained by two dimensional filtering of $\log|S(k,n)|$: (a) formants, (b) plosives, (c) slowly varying component removed. Figure (a) was obtained from Figure 11(c) by clipping below 10% of the maximum value to remove small background features.

CHAPTER 5

TWO DIMENSIONAL COMPRESSION AND EXPANSION

5.1 Review of Homomorphic Compression and Expansion

Oppenheim et al [23] investigated a homomorphic system for compression and expansion of acoustic signals. The system is based on modeling audio signals as a product of two components - a slowly varying, positive envelope signal e(n) and a rapidly varying bipolar signal v(n).

$$s(n) = e(n) \cdot v(n) \tag{5.1}$$

These multiplied signals can be mapped into added signals by the logarithm.

$$\log[s(n)] = \log|e(n)| + \log|v(n)| + i\angle v(n) \tag{5.2}$$

The imaginary part $\angle v(n)$ is either 0 or π, and represents the sign of s(n). If the frequency bands occupied by $\log|e(n)|$ and $\log|v(n)|$ do not overlap, these signals can be separated by linear filtering. Although this will not be the case for intuitive definitions of e(n) and v(n), Reference 23 provides evidence that the component of $\log|s(n)|$ below about 16 Hz should be treated as $\log|e(n)|$.

A multiplicative filter to compress or expand s(n) can be obtained using this model. The logarithm is computed as in 5.2, and the real part is filtered so that $\log|e(n)|$ is multiplied by a

constant p, while $\log|v(n)|$ is passed unchanged. The resulting signal

$$s'(n) = e^{p\hat{}(n)} \cdot v(n) \qquad (5.3)$$

is compressed or expanded as p is less or greater than unity. A block diagram of this system is shown in Figure 13. H(f) is a filter with gain p below 16 Hz and unity gain at higher frequencies.

Figure 14 is a compression-expansion system for transmission through a noisy channel. The blocks labelled "compress" and "expand" are homomorphic systems like that shown in Figure 13. The linear filter used for expansion is the inverse of the filter used for compression. The block "T" represents the effect of channel noise; i.e., audio tape hiss or quantization noise.

The final output of the system, $s'''(n)$, is precisely equal to the input s(n) only in the case where T is the identity system. Small amounts of additive noise, however, do not significantly distort the output. Compression and expansion of the signal makes the signal to noise ratio more uniform over time, and improves performance over that obtained with transmission of the uncompressed signal.

5.2 Two Dimensional Compression and Expansion

In situations where perception of the output signal is important, the performance of the system of Figure 14 may not always be optimum. In speech and music, there are often occasions when the overall signal level is high, but most of the signal energy is concentrated in a relatively small band of frequencies. In such

instances noise at other frequencies may be audible since it is not masked by the input signal.

A compression-expansion system which operates in the time/frequency domain would alleviate this problem by compressing frequency bands more or less independently. Such a system would have the effect of pre-whitening the input signal, which is desireable on a more thecretical basis [18,24].

A two dimensional compression-expansion system can be obtained following the one dimensional homomorphic analysis. The short-time Fourier transform is modeled as a product

$$S(k,n) = E(k,n) \cdot V(k,n) \tag{5.4}$$

where $E(k,n)$ is slowly changing and positive and $V(k,n)$ is rapidly changing and complex. This model has successfully been used as the basis of several speech analysis-synthesis systems (vocoders), and has some physical basis [7]. It leads to the two dimensional multiplicative filter shown in Figure 15, which compresses $S(k,n)$ when $H(f_k, f_n)$ has low frequency gain less than one. A compression-expansion system for the time waveform $s(n)$ can now be constructed by adding the necessary transforms between the time signal and $S(k,n)$. The resulting two dimensional compression-expansion system is shown in Figure 16. (In Figure 16, STFT represents the short-time Fourier transform, 2.7, and $\sum_{k=0}^{N-1}$ represents the reconstruction equation, 2.8.)

The system of Figure 16 is similar to that of Figure 14, but a new theoretical problem exists. Figure 16 is not the identity system, even in the case of a noiseless channel. Synthesis of a

time signal from $S'(k,n)$ followed by reanalysis to produce $S''(k,n)$ is not in general an identity operation. However, as discussed in Chapter 2, $S'(k,n)$ will be approximately equal to $S'_{\cdot}(k,n)$ so long as changes to $S(k,n)$ vary slowly in time (compared to the window), and in frequency (compared to the Fourier transform of the window). This means that the compression and expansion filters, $H(f_n, f_k)$ and $H^{-1}(f_n, f_k)$, can differ from unity only in the low frequency region. This is consistent with the aim of compression and expansion, and accurate reproduction of $\log|S'(k,n)|$ proved not to be a problem in this experiment. The dynamic range of $S(k,n)$ could be compresed 4:1 and reexpanded with an accuracy of 0.3% or better. The worst errors occurred during rapid transients, as would be expected. This accuracy is comparable to that obtained for 4:1 compression and expansion with a homomorphic system like that in Figure 14.

5.3 Comparison of Two Dimensional and Homorphic Compression

The two dimensional system of Figure 16 will not in general lead to the same compressed signal envelope as the homomorphic system of Figure 14. The reason can be understood with the help of a simplified example. Consider a signal which initially has a flat spectrum (e.g., white noise), and assume that at some later time the high frequency half of the signal spectrum has increased by a factor b. We compute the ratio of the final power of the compressed signal to its initial power, and compare the ratio for the two systems.

Assuming that the initial and final signals are locally stationary, the power can be estimated from the square of the signal. In the homomorphic case:

$$P_j \sim \sum_{l=-L}^{L} |e^p(j)\cdot v(j+l)|^2 = e_j^{2p}\cdot\sum_{l=-L}^{L} |v(j+l)|^2, \tag{5.5}$$

$$R_1 = P_f/P_i = (e_f^2/e_i^2)^p, \tag{5.6}$$

$$R_1 = (1/2 + b/2)^p. \tag{5.7}$$

Equation 5.7 was obtained using Parseval's theorem and the known signal spectrum. In the two dimensional case:

$$P_j \sim \sum_{l=-L}^{L} |s(j+l)|^2, \tag{5.8}$$

$$\sim \sum_{l=-L}^{L} |\sum_{k=0}^{N-1} E^p(k,j)\cdot V(k,j+l)\cdot\exp[i\Omega_k(j+l)]|^2, \tag{5.9}$$

$$\sim \sum_{l=-L}^{L}\sum_{k=0}^{N-1} |E^p(k,j)\cdot V(k,j+K)|^2, \tag{5.10}$$

$$P_j \sim \sum_{k=0}^{N-1} |E^2(k,j)|^p \cdot \sum_{l=-L}^{L} |V(k,j+K)|^2. \tag{5.11}$$

The absolute square can be taken inside the sum in 5.10 because $E^p(k,j)\cdot V(k,j)\cdot\exp(i\Omega_k j)$ are approximately nonoverlapping band pass signals, and therefore the cross products in 5.9 sum to zero when time averaged. Assuming that the time average behavior of V(k,n) is the same for all k, we obtain the desired ratio.

$$R_2 = P_f/P_i = (\sum_{k=0}^{N-1} |E^2(k,f)|^p)/(\sum_{k=0}^{N-1} |E^2(k,i)|^p) \tag{5.12}$$

$$R_2 = 1/2 + b^p/2 \tag{5.13}$$

Figure 17 is a plot of the ratio R_2/R_1 for the example just described, and for the case where 1/4 of the spectrum increases. The curves were obtained with p = 0.25, but are not a sensitive function of p. They show what might be expected on a physical basis: the two dimensional system compresses the signal more when the energy is concentrated in a narrow frequency band. This effect was observed experimentally.

5.4 Experimental Results

Two experiments were performed with the two dimensional system of Figure 16. The experiments were intended to simulate transmission of speech through an analog and a digital channel. In both cases, the results were judged by comparison with a similar homomorphic system.

A 20.0 msec hanning window was used for the short-time Fourier transform to provide maximum frequency resolution consistent with time direction compression below about 16 Hz. The linear filter in the compressor had a unit sample response equivalent to the continuous function

$$h(k,n) = \delta(k,r) - 0.75 \cdot w(k,n) \tag{5.14}$$

$$w(k,n) = A \cdot [.5+.5 \cdot \cos(\pi k/390.)] \cdot [.5+.5 \cdot \cos(\pi n/.04)] \tag{5.15}$$

$$-.04 < n < .04 \quad ; \quad -390. < k < 390.$$

where t is in seconds and K is in Hz. A is a normalizing factor so that the area under w(k,n) is unity. This resulted in a high pass filter with transitions at about $f_n = 13$ Hz, $f_k = 1.3$ msec, and a low frequency gain of 0.25. The impulse response was specified rather than the frequency response because a smooth impulse response was essential to accurate expansion. The filter in the homomorphic system used for comparison was

$$h(n) = \delta(n) - 0.75 \cdot B \cdot [.5+.5 \cdot \cos(\pi n/.04)] \tag{5.16}$$

where B is again a normalizing factor.

An analog channel was simulated by adding gaussian distributed white noise to the compressed speech, s'(n), at levels from -36 dB

to -12 dB. The level of the noise was referenced to the peak compressed signal energy determined by averaging the squared signal with a 50 msec hanning window. The peak power occurred at the same place (during a vowel sound) for both homomorphic and two dimensional compression, and the peak power criterion did not appear to bias the results in favor of either method. Figures 18 and 19 show the compressed speech s'(n), compressed speech plus noise s''(n), and the expanded output speech s'''(n) for the homomorphic system (Figure 14) and the two dimensional system (Figure 16). Figure 20 shows a comparison of the output speech for the two systems and an example where no compression was used. Figure 21 shows spectrograms of the output. (The noise above 4000 Hz in Figure 21(b) was removed by the anti-imaging filter on playback.) All four figures are for a channel signal to noise ratio of 12 dB.

The two dimensional system provided 18 dB improvement over the homomorphic system. Noise was first audible in the output of the two dimensional system at a channel signal to noise ratio of 12 dB, compared to 30 dB for the homomorphic system. Some of the improvement can be attributed to the removal of the natural high frequency rolloff of speech by compression of S(k,n), as evidenced in Figure 21. However, flattening the average speech spectrum by preemphasis at 6 dB/oct above 400 Hz prior to compression and complementary deemphasis of the output signal improved the performance of the homomorphic system by only ~3 dB.

A digital channel was simulated by quantizing the compressed signal to a small number of bits (3-8). The quantizer is shown in Figure 22. Since envelope distortion caused by clipping is

particularly noticeable in both systems, the compressed speech was scaled so that the peak signal was equal to 2^{n-1}, where n is the number of bits. A dithering signal was added prior to quantization and subtracted afterward to break up the correlation between the quantization noise and the signal [25,26]. The dithering signal used was uniformly distributed pseudorandom noise, with peak magnitude equal to 1/2 quantization level and zero mean.

The results of this experiment are consistent with the analog channel simulation. The two dimensional system provided a three bit improvement over the homomorphic system. Noise was first audible in the output of the two dimensional system with four bit channel quantization, compared to seven bit quantization for the homomorphic system. For comparison, quantization noise was first audible in the uncompressed signal at nine bit quantization. (Fifteen bit quantization was used for the input and output speech.) The apparent loss of 6 dB signal to noise ratio compared to the analog experiment, based on the "6 dB per bit" rule of thumb, is due to the difference in criteria for scaling the compressed signal.

The waveforms for three bit quantization of the channel signal are nearly indistinguishable from Figures 18 - 21. The same is true of listening tests.

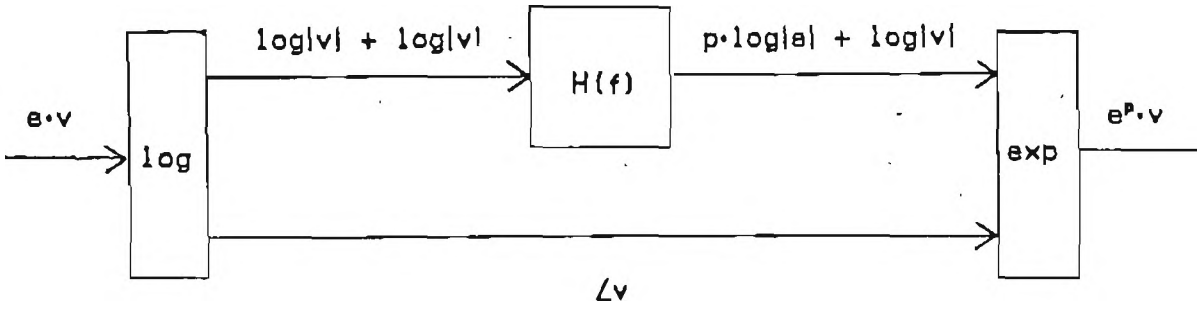


FIGURE 13

Homomorphic filter for compression or expansion of acoustic signals.

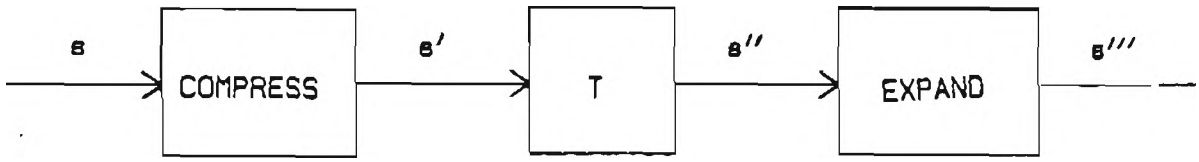


FIGURE 14

Homomorphic compression/expansion system for transmitting acoustic signals through a noisy channel.

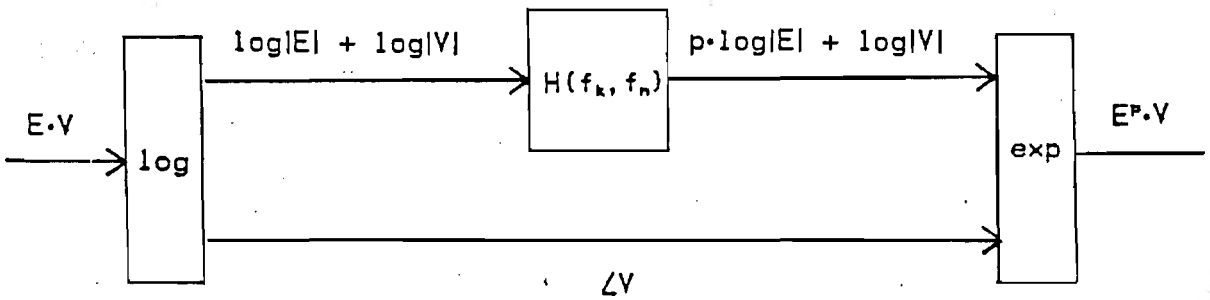


FIGURE 15

Two dimensional homomorphic filter for compression or expansion of $|S(k,n)|$.

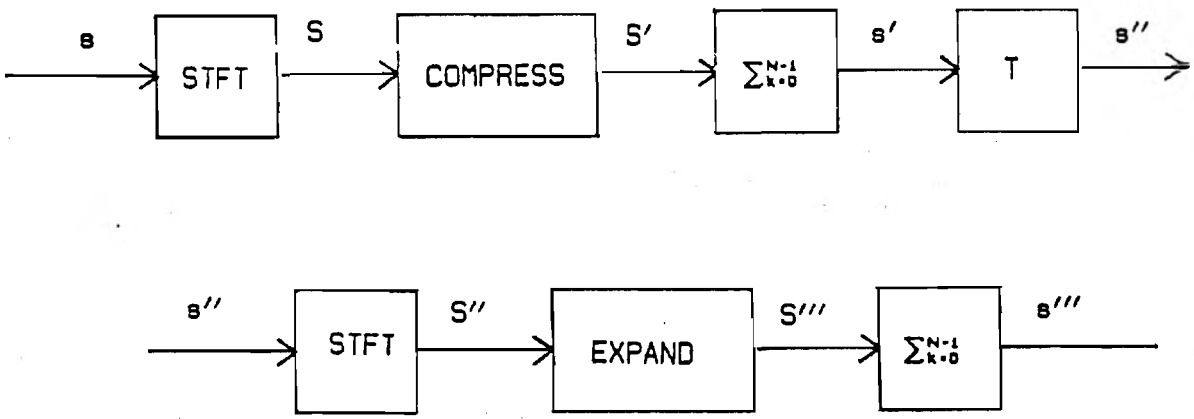


FIGURE 16

Two dimensional compression/expansion system for transmitting acoustic signals through a noisy channel.

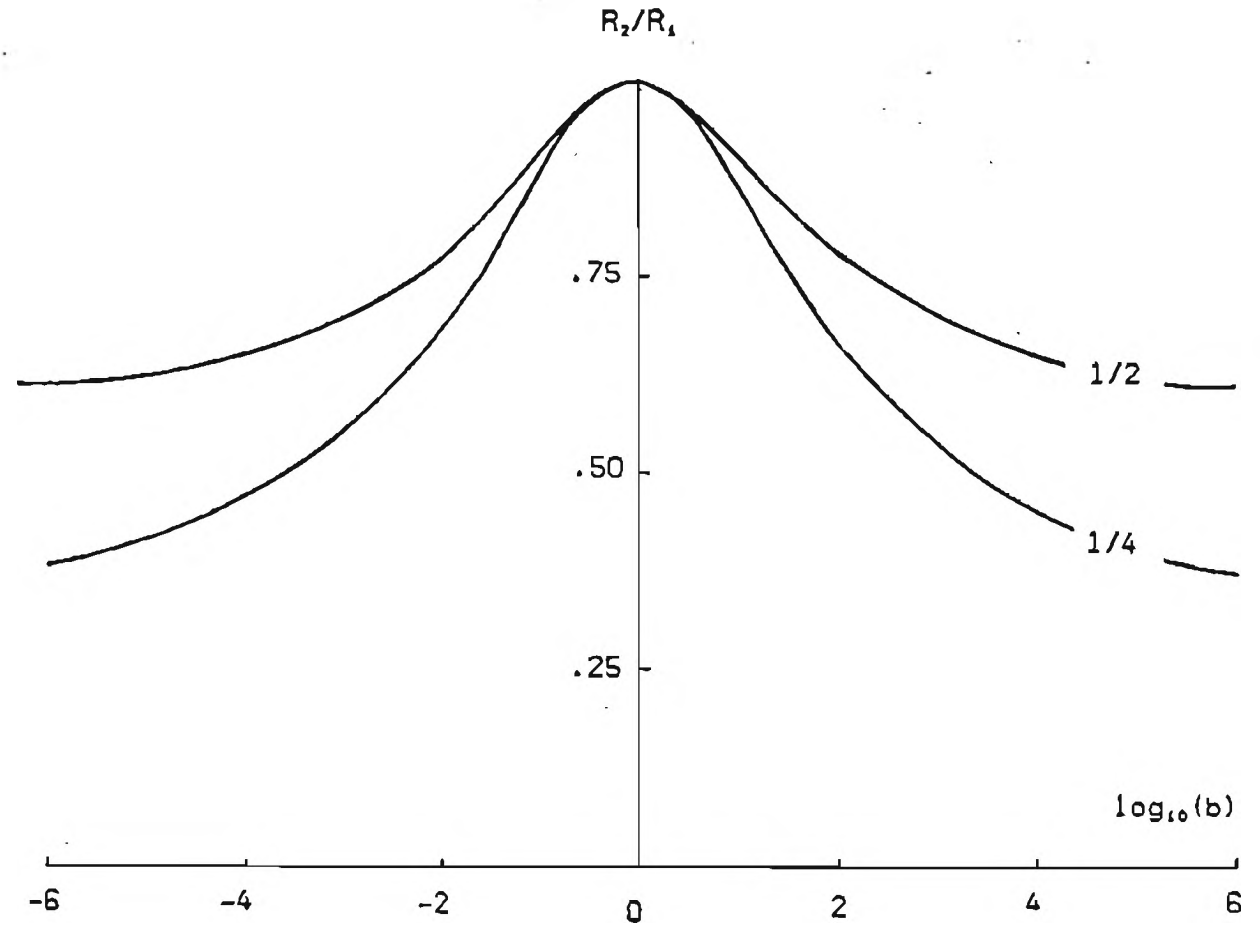
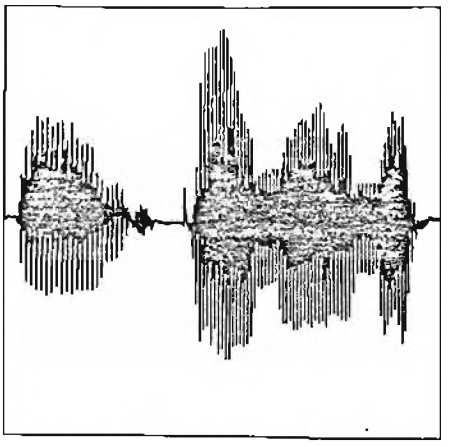
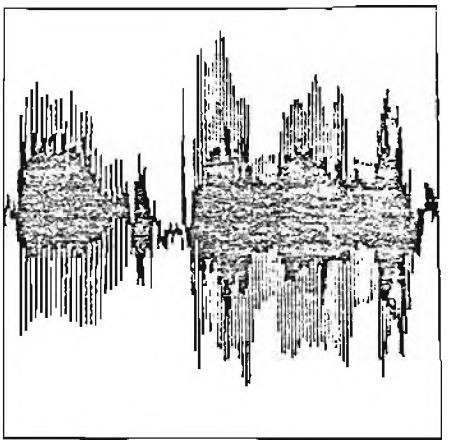


FIGURE 17

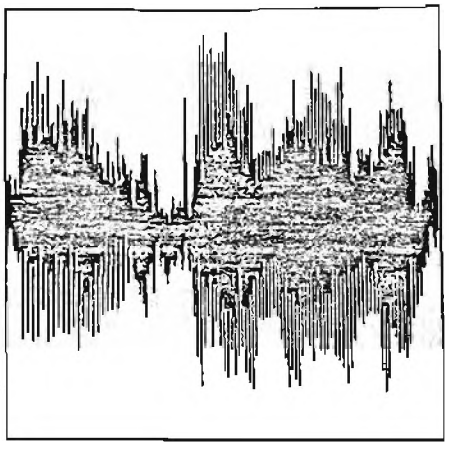
Comparison of homomorphic and two dimensional compression. The curves show the ratio of two dimensional compression to homomorphic compression, R_2/R_1, as a function of the change in a portion of the spectrum of the input signal, b. The two curves represent changes in 1/2 and 1/4 of the spectrum, as labelled.



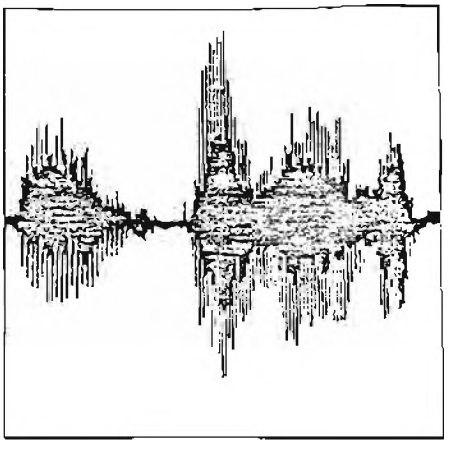
(a)



(b)



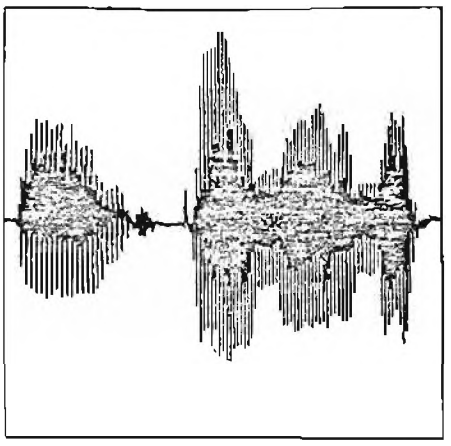
(c)



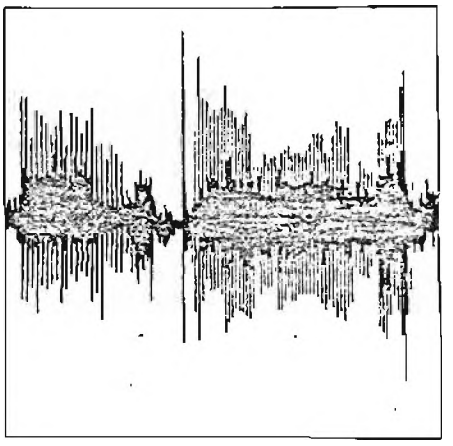
(d)

FIGURE 18

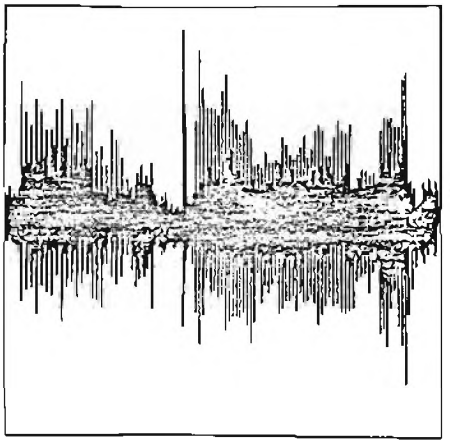
Compression and expansion of speech using a homomorphic system: (a) original speech, (b) compressed speech, (c) compressed speech with noise added, (d) expanded output speech. Each picture represents one second of speech. Channel signal to noise ratio is 12 dB.



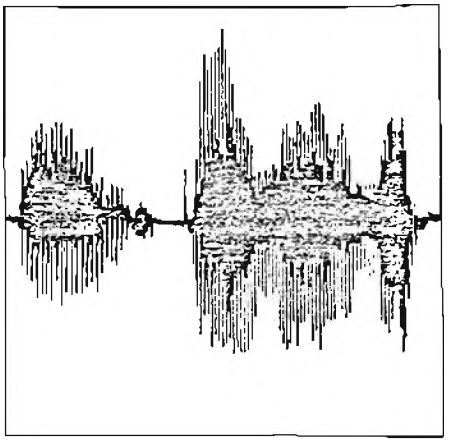
(a)



(b)



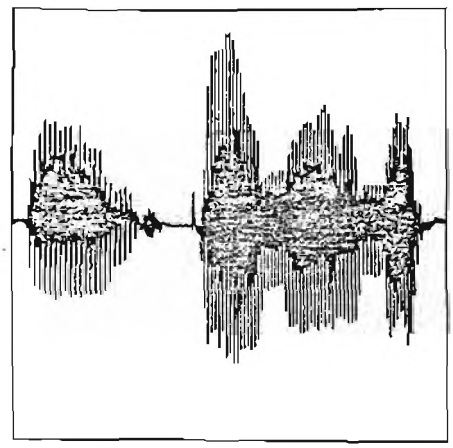
(c)



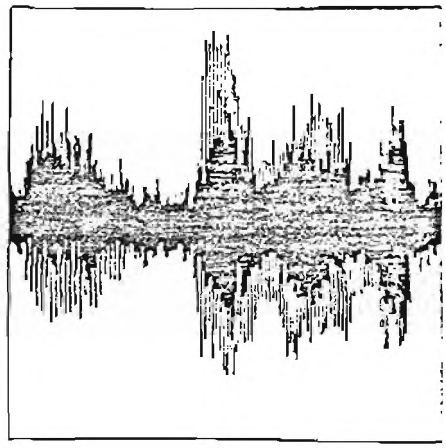
(d)

FIGURE 19

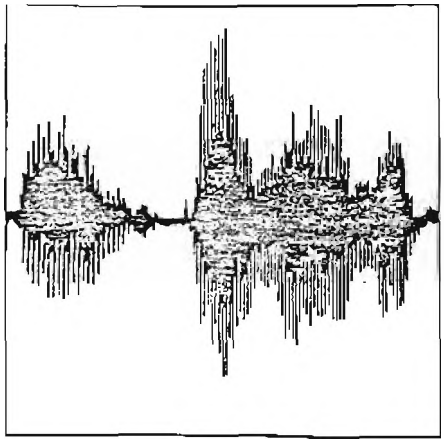
Compression and expansion of speech using a two dimensional system: (a) input speech, (b) compressed speech, (c) compressed speech with noise added, (d) expanded output speech. Each picture represents one second of speech. Channel signal to noise ratio is 12 dB.



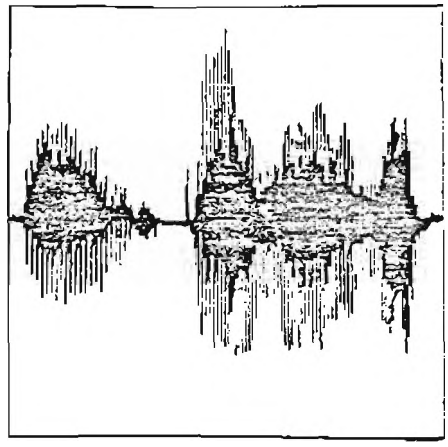
(a)



(b)



(c)

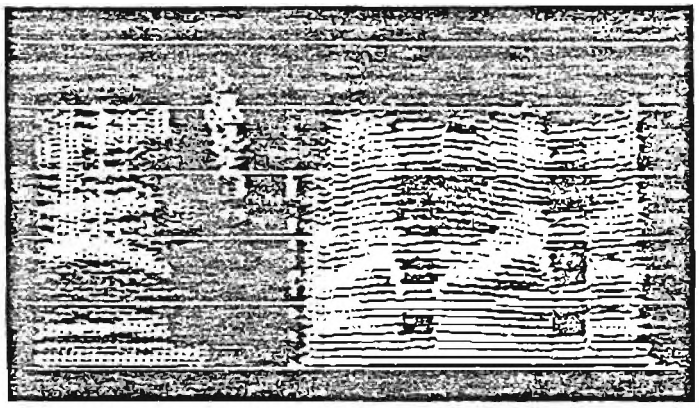


(d)

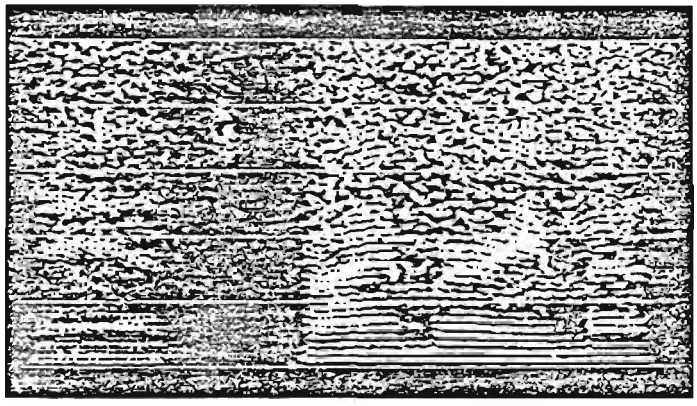
FIGURE 20

Comparison of output speech for homomorphic and two dimensional systems: (a) input speech, (b) no compression, (c) homomorphic system, (d) two dimensional system. Each picture represents one second of speech. Channel signal to noise ratio is 12 dB.

(a)



(b)



(c)

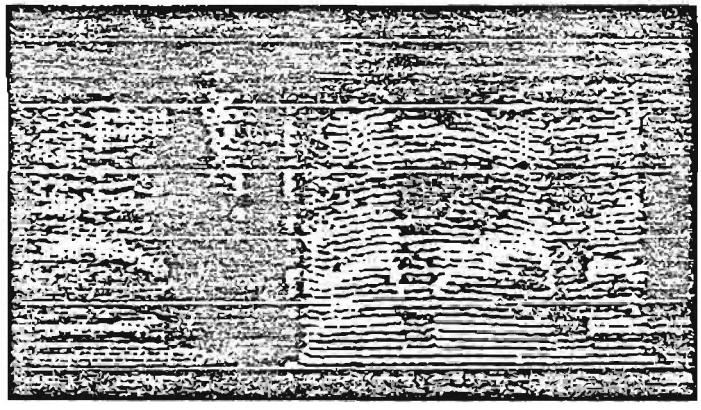


FIGURE 21

Spectrograms of output speech for homomorphic and two dimensional systems: (a) input speech, (b) homomorphic system, (c) two dimensional system. Channel signal to noise ratio 12 dB. All three spectrograms have been scaled 6 dB/oct above 400 Hz.

CHAPTER 6

REMOVAL OF LOCALLY PERIODIC INTERFERING SIGNALS

6.1 Removal by Spectrum Estimation

Cases arise where speech intelligibility is degraded by locally periodic signals. Electrical interference in communications channels is sometimes periodic with many harmonics. Aircraft cockpit noise often has a large periodic component due to engine noise. The precise fundamental frequency and harmonic structure of these signals is typically hard to predict, and the signals change from time to time. Electrical interference may be intermittant or fade, and cockpit noise changes with engine speed. For times of the order of seconds, however, these signals appear periodic.

The short-time spectrum of periodic signals is characterized by sharply defined harmonic structure. An adaptive filter which selectively attenuates such features can be obtained using a technique similar to that developed by Stockham et al. [27] for blind deconvolution, in which the actual spectrum of the noisy signal is replaced with a prototype spectrum obtained from undegraded speech. The advantage of this approach is that precise knowledge of the frequency and harmonic structure of the interfering signal is not required.

The local spectrum of the degraded speech can be estimated by time averaging the short-time spectrum

$$\Phi(k,n) \simeq (2L+1)^{-1} \sum_{l=-L}^{L} |S(k,n+l)|^2 \qquad (6.1)$$

where the length of the time average is determined by the stationary time of the interfering signal. The estimated spectrum in 6.1 will have two components, one due to the interfering signal and another due to the local average speech spectrum. If the averaging time is long enough, the local average speech spectrum can be approximated by a prototype computed from similar undegraded speech. Portions of the local spectrum which are dominated by noise can then be restored by multiplying the short-time Fourier transform by

$$H(k,n) = [P(k)/\Phi(k,n)]^{1/2} \qquad (6.2)$$

where P(k) is the prototype spectrum.

Equation 6.2 is the square root of the Wiener filter which would be obtained if we modeled speech as a stationary random process, uncorrelated with the interfering signal. It might therefore be expected that 6.2 will not adequately attenuate frequencies dominated by the interferance. This is also to be expected because 6.2 only restores the magnitude of the short-time transform, and portions of the transform which were dominated by noise will still sound like noise due to the phase. However, squaring 6.2 may not be advisable, because this will emphasize distortions caused by the fact that P(k) is not precisely the local average speech spectrum. Equation 6.2 is therefore replaced with

$$H(k,n) = [P(k)/\Phi(k,n)]^{p} \qquad (6.3)$$

where p is between 0.5 and 1. It was found experimentally that for

averaging times of the order of a second $p = 0.75$ represents a good compromise between noise reduction and distortion of the underlying speech.

6.2 Removal by Two Dimensional Filtering

The short-time spectrum can alternately be restored by two dimensional filtering of $\log\{|S(k,n)|^2\}$. This approach does not require a prototype speech spectrum, and so may be useful when the stationary time of the interfering signal is relatively short.

This approach is based on the fact that low pass filtering of $\log\{|S(k,n)|^2\}$ in the time direction is equivalent to estimating $\log\{\Phi(k,n)\}$ by a weighted time average of $\log\{|S(k,n)|^2\}$. Spectrum estimates based on time averages of the logarithm of the local spectrum are biased toward stationary spectral components [27]. In the present application, this bias is an advantage since it emphasizes the spectrum of the interfering signal.

Time direction low pass filtering attenuates all points in the (two dimensional) frequency domain except those which are on or near the f_k axis. If the interfering signal has appreciable harmonic structure, features in the Fourier transform of $\log\{|S(k,n)|^2\}$ due to the interfering signal lie mainly beyond $|f_k| = \tau$, where τ is the period of the interfering signal. Features for $|f_k| < \tau$ are primarily due to the local speech spectrum. The spectrum of the interfering signal can therefore be estimated with a filter $L(f_n, f_k)$ which isolates components near the f_k axis for $|f_k| > \tau$. Such a filter is shown in Figure 23. This filter can be implemented quite efficiently, since its frequency response (and impulse response) are separable.

The magnitude of the short-time spectrum can be restored by filtering $\log\{|S(k,n)|^2\}$ with

$$H(f_n, f_k) = 1 - p \cdot L(f_n, f_k). \qquad (6.4)$$

where p is a number between 0.5 and 1 analogous to the power in 6.3.

6.3 Experimental Results

To evaluate the processes discussed above, a test signal was constructed by adding an assymetric square wave (positive signal 1.5 times as long as negative signal) to speech. The frequency of the square wave was changed from time to time. The approximate signal to noise ratio was -26 dB, and the speech was unintelligible.

The short-time Fourier transform of the degraded speech was calculated with a 102.4 msec hanning window. A longer window (~200 msec) would have better resolved the interfering signal, but the length of the window was limited by computer memory requirements in the implementation of the short-time transform used here. $|S(k,n)|$ was modified using 6.3 with $p = 0.75$, and an averaging time of one second for $\Phi(k,n)$. The prototype was obtained from a different speaker.

The time waveform and a spectrogram of the degraded and restored signals are shown in Figure 25. The interfering signal is greatly attenuated, except near places where the frequency of the interfering signal changes. The resulting speech is intelligible, and is sufficiently natural that the speaker is recognizable.

$|S(k,n)|$ was also restored by filtering $\log\{|S(k,n)|^2\}$ as in 6.4, with $p = 0.75$ and a cutoff frequency $f_n = 0.5$ Hz. In this case

the noise was almost entirely removed, except near frequency changes, and the resulting speech was quite intelligible. The speech was more natural and less noisy than that obtained by spectrum averaging. The processed signal is shown in Figure 25.

Filtering of $\log\{|S(k,n)|^2\}$ was also applied to removal of electrical interference from several tape recorded signals obtained in more realistic circumstances. The interfering signals were almost completely removed in each case. However, the intelligibility of the processed speech was sometimes poor due to other distortions, such as resonances in the recording system and tape saturation.

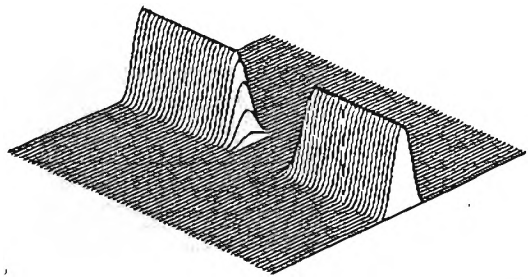


FIGURE 23

Frequency response of filter for removal of periodic interfering signals. The origin is at the center. The passbands are along the f_k axis for $f_t > \tau$. (The picture is not to scale.)

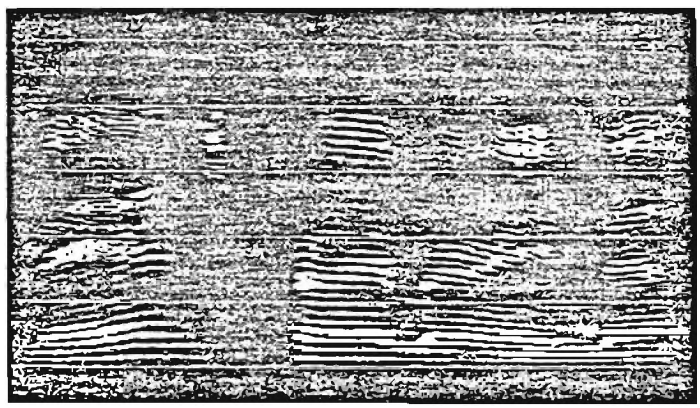
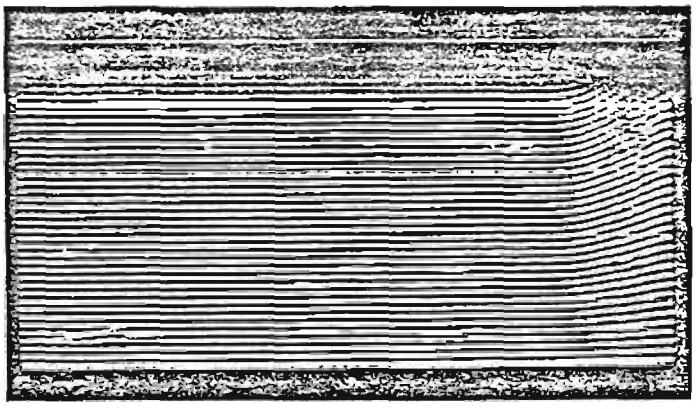


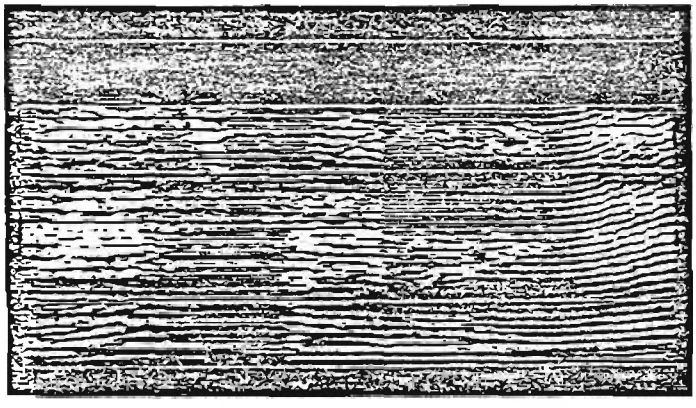
FIGURE 24

Original (undegraded) speech used in experiment on removal of locally periodic interference shown in Figure 25.

(a)



(b)



(c)

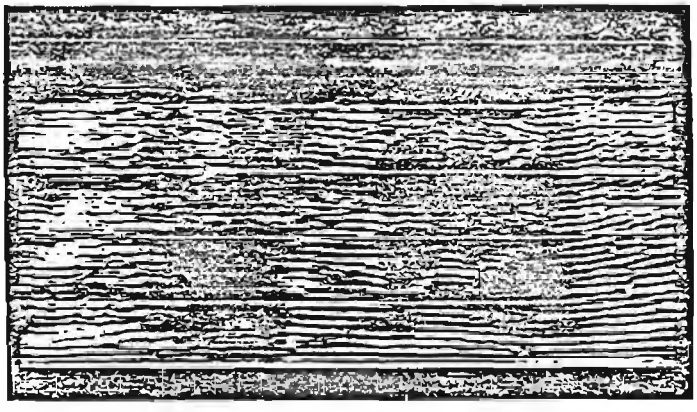


FIGURE 25

Removal of locally periodic interference: (a) speech degraded with square wave noise, (b) restored by spectrum averaging (c) restored by two dimensional filtering. All three spectrograms have been scaled 6 dB/oct above 400 Hz.

CHAPTER 7

CONCLUSIONS

The experiments described here indicate that acoustic signal processing in the time/frequency domain can offer significant advantages over conventional one dimensional methods, particularly when perception is the final measure of system performance. These advantages are gained at the expense of computational complexity, but the computing hardware exists (e.g., fast array processors) to make time/frequency processing practical when the problem merits. The chief disadvantage of time/frequency processing is the paucity of concise mathematical results describing the effect of such processing.

All of the experiments described here were based on the magnitude of the short-time Fourier transform. Phase information was used only to reconstruct a time signal. In at least two cases, the results might be improved using phase information.

Removal of broad band background noise by thresholding as described in Chapter 3 attenuates signal harmonics when their magnitude remains too near the noise level. It is possible that these harmonics could be detected in the phase. Consider $\angle S(k,n)$ as a function of time. Noise will have phase randomly distributed between $-\pi$ and π, while coherent signals such as harmonics will have nearly linear phase. The phase time derivitive, therefore, should

have appreciable high frequency components only when the short-time transform is locally dominated by noise. Techniques for detection of FM signals in noise might even be applied to separate out the component of the phase due to the signal and reject that due to noise.

Locally periodic interference like that discussed in Chapter 6 might be rejected on similar criteria. In this case, however, the portions of S(k,n) dominated by the interfering signal will have linear phase while those due mainly to speech will be more random. Attenuation of the interfering signal based on phase criteria might produce less distortion in the underlying speech than the methods of Chapter 6.

It was mentioned in the Introduction that the effectiveness of the short-time Fourier transform as a perceptual model was limited because it is constant bandwidth, while the ear is more nearly constant Q. The potential advantages of a constant Q transform were also observed experimentally. For example, in background noise removal, narrower bandwiths could be used to better resolve low frequency components, since the low frequency harmonics change slowly with time. The same is true for analysis of any signal with harmonic structure - increasing the bandwidth with frequency would provide more nearly optimum resolution of the entire spectrum, since high frequency harmonics change more rapidly. A reasonably efficient algorithm for the constant Q equivalent of the short-time transform would certainly improve some of the results obtained here, and would make possible detailed studies of the perception of acoustic stimuli similar to those which have been done for the

visual system [28].

One application of time/frequency processing not considered here which appears especially promising is hearing aids for the deaf. A large number of persons with impaired hearing suffer from recruitment [29], a frequency dependent reduction in the dynamic range of the auditory system. Recruitment is usually due to damage to the hair cells or auditory nerve, and raises the threshold of hearing in the region affected. The threshold of discomfort is not changed, however, so that it is often not practical to preemphasize the signal sufficiently to compensate for the hearing loss. The results in Chapter 4 indicate that frequency dependent compression based on the changes in threshold of hearing might be more effective, particularly in improving speech intelligibility.

APPENDIX A

A LIMIT ON THE UNCERTAINTY PRODUCT FOR THE SHORT-TIME SPECTRUM

In this appendix, a relation is derived for the short-time spectrum which is similar to the two dimensional uncertainty principle in quantum mechanics. This result shows how the usual uncertainty relation for signals and window functions carries over to the short-time spectrum. We define

$$\Delta^2 t = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t-t_0)^2 \cdot |S(\omega,t)|^2 \, d\omega \, dt}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |S(\omega,t)|^2 \, d\omega \, dt} , \tag{A.1}$$

$$\Delta^2 \omega = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\omega-\omega_0)^2 \cdot |S(\omega,t)|^2 \, d\omega \, dt}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |S(\omega,t)|^2 \, d\omega \, dt} . \tag{A.2}$$

It will be shown in this appendix that

$$\Delta\omega \Delta t \geq 1. \tag{A.3}$$

The short-time fourier transform is defined [1]

$$S(\omega,t) = \int_{-\infty}^{\infty} f(x) \cdot m(t-x) \cdot \exp(-i\omega x) \, dx \tag{A.4}$$

$$= (2\pi)^{-1} \int_{-\infty}^{\infty} F(\lambda+\omega) \cdot M(\lambda) \cdot \exp(i\lambda t) \, d\lambda . \tag{A.5}$$

The short-time spectrum is the squared magnitude of the short-time Fourier transform. Equation A.5 is obtained by viewing A.4 as a convolution, and using the convolution theorem to rewrite it as the Fourier transform of a product. $F(\omega)$ and $M(\omega)$ are the Fourier transforms of $f(t)$ and $m(t)$.

First, the denominator of A.1 and A.2 are evaluated using Parseval's theorem and the above definition. We will use the notation $\|g\| = \int_{-\infty}^{\infty} |g(\rho)|^2 \, d\rho$.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |S(\omega,t)|^2 \, d\omega \, dt$$

$$= 2\pi \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(x)|^2 \cdot |m(t-x)|^2 \, dx \, dt$$

$$= 2\pi \|f\| \cdot \|m\| \tag{A.6}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |S(\omega,t)|^2 \, d\omega \, dt = (2\pi)^{-1} \|F\| \cdot \|M\| \tag{A.7}$$

We now evaluate ω_o and t_o, again using Parseval's theorem. For simplicity, it is assumed that the window is real and centered in time.

$$t_o = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t \cdot |S(\omega,t)|^2 \, d\omega \, dt}{2\pi \|f\| \cdot \|m\|} \tag{A.8}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t \cdot |f(x)|^2 \cdot |m(t-x)|^2 \, dx \, dt}{\|f\| \cdot \|m\|} \tag{A.9}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t+x) \cdot |f(x)|^2 \cdot |m(t)|^2 \, dx \, dt}{\|f\| \cdot \|m\|} \tag{A.10}$$

$$= \frac{\int_{-\infty}^{\infty} x \cdot |f(x)|^2 \, dx}{\|f\|} \tag{A.11}$$

The first term in A.10 integrates to zero because the window is centered.

$$\omega_o = \frac{2\pi \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \omega \cdot |S(\omega,t)|^2 \, d\omega \, dt}{\|F\| \cdot \|M\|} \tag{A.12}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \omega \cdot |F(\lambda+\omega)|^2 \cdot |M(\lambda)|^2 \, d\omega \, d\lambda}{\|F\| \cdot \|M\|} \tag{A.13}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\omega-\lambda) \cdot |F(\omega)|^2 \cdot |M(\lambda)|^2 \, d\omega \, d\lambda}{\|F\| \cdot \|M\|} \tag{A.14}$$

$$= \frac{\int_{-\infty}^{\infty} \omega \cdot |F(\omega)|^2 \, d\omega}{\|F\|} \tag{A.15}$$

We are now in a position to evaluate $\Delta^2 t$, $\Delta^2 \omega$ in terms of the signal and the window.

$$\Delta^2 t = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t-t_0)^2 \cdot |f(x)|^2 \cdot |m(t-x)|^2 \, dx \, dt}{\|f\| \cdot \|m\|} \tag{A.16}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t+x-t_0)^2 \cdot |f(x)|^2 \cdot |m(t)|^2 \, dx \, dt}{\|f\| \cdot \|m\|} \tag{A.17}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [t^2+(x-t_0)^2] \cdot |f(x)|^2 \cdot |m(t)|^2 \, dx \, dt}{\|f\| \cdot \|m\|} \tag{A.18}$$

$$= \frac{\|(t-t_0)\cdot f\|}{\|f\|} + \frac{\|tm\|}{\|m\|} \tag{A.19}$$

The cross term from A.17 integrates to zero - see A.11.

$$\Delta^2 \omega = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\omega-\omega_0)^2 \cdot |F(\lambda+\omega)|^2 \cdot |M(\lambda)|^2 \, d\omega \, d\lambda}{\|F\| \cdot \|M\|} \tag{A.20}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [(\omega-\omega_0)^2+\lambda^2] \cdot |F(\omega)|^2 \cdot |M(\lambda)|^2}{\|F\| \cdot \|M\|} \tag{A.21}$$

$$= \frac{\|(\omega-\omega_0)F\|}{\|F\|} + \frac{\|\omega M\|}{\|M\|} \tag{A.22}$$

Introducing the notation

$$\Delta^2 g = \frac{\|(\rho-\rho_0)\,g\|}{\|g\|} \, ,$$

we can summarize the above results in a concise form.

$$\Delta^2 t = \Delta^2 f + \Delta^2 m \tag{A.23}$$

$$\Delta^2 \omega = \Delta^2 F + \Delta^2 M \tag{A.24}$$

The fact that the time and frequency uncertainty of the short-time spectrum can be decomposed into the sum of the corresponding uncertainty for the signal and the window shows in a simple manner the effect of the window on the short-time spectrum.

The usual objective in short-time frequency analysis is not to minimize the product $\Delta^2\omega\,\Delta^2 t$, but to minimize the portion due to the window. Equations A.23 and A.24 suggest the use of smooth window functions like the gaussian which minimize $\Delta^2 m\,\Delta^2 M$.

Inspection of A.18 and A.21 show that minimizing the product $\Delta^2\omega\,\Delta^2 t$ is equivalent to minimizing the uncertainty product for the two dimensional, separable function $f(x)\cdot m(t)$ and its Fourier transform. The result is a well known property of the two dimensional Fourier transform, most often encountered in quantum mechanics [16].

$$(\Delta^2 f+\Delta^2 m)\cdot(\Delta^2 F+\Delta^2 M) \geq 1 \qquad \text{(A.25)}$$

The equality in A.25 is satisfied when the signal and the window are gaussians with the same variance.

$$f(t) = A\cdot\exp[-(t-t_0)^2/2\sigma^2]\cdot\exp(i\omega_0 t)$$

$$m(t) = B\cdot\exp(-t^2/2\sigma^2)$$

APPENDIX B

EXPERIMENTAL METHODS

B.1 Computational Information

The experiments described in this paper were performed using a general purpose Digital Equipment Corporation PDP-10 computer, with 65,536 36-bit words of memory. The main programs were written in FORTRAN, but made use of some assembly language (MACRO-10) subroutines to speed up repetitive operations. All calculations were done with floating point arithmetic. Magnetic disk storage was used for the short-time Fourier transform and other arrays too large to fit in memory.

No explicit effort was made to optimize the computational effeciency of the algorithms used, as the main objective of this research was to demonstrate the capabilities of time/frequency processing. However, the computing times required do provide an estimate of the computational effort inherent in this type of process.

The times for floating point addition and multiplication on the PDP-10 are 5 and 11 microseconds, respectively. A quite efficient assembly language subroutine for the Fast Fourier Transform was available, which required $42 \cdot K \log_2(K)$ microseconds to compute the discrete Fourier transform of a K point complex sequence (e.g., 2048 point sequence requires ~1 sec). This subroutine was used with a

FORTRAN implementation of Portnoff's method [15] for calculating the discrete short-time Fourier transform and its inverse.

The short-time transform with a hanning window was sampled in time at a rate of 8 samples per window length. This resulted in a computing time of about 1 minute for the short-time transformin polar form, and inverse transform of 10,000 samples (one second) of signal. The time varies slightly (see equation 2.14) dependant on the number of frequency samples in the short-time transform. The transform time is increased when the window is augmented with zeros to minimize aliasing, as for two dimensional compression and expansion in Chapter 5.

The local Wiener filtering and thresholding described in Chapter 3 required 30 seconds to 1 minute per second of signal, in addition to the transform time. The time for two dimensional filtering varied widely, depending on the specific filter used. The best times occurred when the filter impulse response was separable and of realtively small extent, so that convolution could be implemented directly. In these cases, filtering time was about 2 minutes per second of signal.

B.2 Recording and Playback of Signals

The speech used in this research was digitized and stored on magnetic disks in real time, using a B and K model 4144 condenser microphone, low noise amplifiers, and a 15 bit analog to digital converter. The speech was prefiltered at 4,000 Hz, and sampled at 10,000 samples per second. The recordings were made in a sound isolated but acoustically live room. The overall signal to noise ratio for the electronics at the time of recording was 80 dB.

The Caruso recording used in Chapter 3 was previously digitized by T. G. Stockham, Jr.

Digitized signals were played back through a 16 bit digital to analog converter, and filtered at 4,000 Hz. Critical listening was done with Koss PRO-4A headphones. Signals could also be reproduced with Bose 901 loudspeakers.

B.3 Spectrogram Displays

The spectrograms in this paper were produced with a precision cathode ray tube display. The output signal is compensated to account for the properties of the cathode ray tube phosphor and the photographic film, so that the intensity of the light reflected from the picture is proportional to the signal in the computer.

The spectrograms are 256 X 512 point displays of the magnitude of the short-time Fourier transform. The magnitude is scaled to use the full dynamic range of the film; in some cases, clipping of the brightest features is allowed in order to improve the visibility of fainter regions. The magnitude is normally scaled by 6 dB/oct above 400 Hz to remove the natural high frequency rolloff of speech.

With the exception of Chapter 4, all spectrograms in this paper were made by reanalyzing the processed signal using the same window as the process. Those in Chapter 4 show the modified short-time spectrum directly, since a modified signal was not synthesized.

The speech shown in the spectrograms was that of a male with relatively low pitch. For this reason, the spectrograms sometimes show intermediate resolution of pitch where temporal structure and harmonics are simultaneously visible. This is evident in the word

"man's" in Figure 5(b), for example. This had no significant effect on processing since only low frequency components of the short-time spectrum were modified.

REFERENCES

[1] C. J. Weinstein, "Short-Time Fourier Analysis and Its Inverse," M.S. Thesis, Elec. Eng. Dep., Mass. Inst. Technol., 1966.

[2] J. L. Flanagan, Speech Analysis, Synthesis, and Perception. 2nd ed., New York: Springer-Verlag, 1972.

[3] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," IEEE Spectrum, vol 7, pp 57-62, August 1970.

[4] A. M. Liberman, P. C. Delattre, and F. S. Cooper, "The role of selected stimulus-variables in the perception of the unvoiced stop consonants," Amer. J. Psychol., vol 65, pp 1-13, October 1952.

[5] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychol. Monographs, vol 68, no 379, 1954.

[6] J. L. Flanagan and R. M. Golden, "Phase vocoder," Bell Syst. Tech. J., vol 45, pp 1493-1509, November 1966.

[7] M. R. Schroeder, "Vocoders: analysis and synthesis of speech, a review of 30 years of applied research," Proc. IEEE, vol 54, pp 720-734, May 1966.

[8] H. L. v. Helmholz, On the Sensations of Tone. Translation of the fourth German edition of 1877 by A. J. Ellis, New York: Dover, 1954.

[9] J. V. Tobias, Foundations of Modern Auditory Theory, New York: Academic Press, 1970.

[10] M. R. Schroeder, "Models of hearing," Proc. IEEE, vol 63, pp 1332-151, September 1975.

[11] G. v. Bekesy, Experiments in Hearing. New York: McGraw-Hill, 1960.

[12] N. Kiang, Discharge Patterns of Single Fibres in the Cat's Auditory Nerve. Cambridge, Mass: M.I.T. Press, 1965.

[13] A. V. Oppenheim and R. W. Schafer, Digital Signal Processing. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[14] E. A. Guillemin, Theory of Linear Physical Systems. New York: Wiley, 1963.

[15] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," to be published in IEEE Trans. Acous., Speech, Signal Processing.

[16] K. Gottfried, Quantum Mechanics. New York: Benjamin, 1966: vol 1, p 214.

[17] H. J. Landau and H. O. Pollack, "Prolate spheroidal wave functions, Fourier analysis and uncertainty - II," Bell Syst. Tech. J., vol 40, pp 65-85, January 1961.

[18] R. B .Blackman and J. W. Tukey, The Measurement of Power Spectra. New York: Dover, 1968.

[19] H. Y. Huang, "A Collection of Digital Window Functions," M.S. Thesis, Comput. Sci. Dep., Univ. Utah, Salt Lake City, 1973.

[20] N. Weiner, The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. New York: Wiley, 1949.

[21] M. B. Sachs and N. Kiang, "Two-tone inhibition in auditory nerve fibers," J. Acoust. Soc. Amer., vol 43, pp 1120-1128, May 1968.

[22] T. G. Stockham, Jr., "Image processing in the context of a visual model," Proc. IEEE, pp 828-842, July 1972.

[23] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, Jr., "Nonlinear filtering of multiplied and convolved signals," Proc. IEEE, vol 56, pp 1264-1291, August 1968.

[24] J. P. Costas, "Coding with linear systems," Proc IRE., vol 40, pp 1101-1103, September 1952.

[25] N. S. Jayant and L. R. Rabiner, "The application of dither to the quantization of speech signals," Bell Syst. Tech. J. , vol 51, pp 1293-1304, July-August 1972.

[26] L. R. Rabiner and J. A. Johnson, "Perceptual evaluation of the effects of dither on low bit rate pcm systems," Bell Syst. Tech. J., vol 51, pp 1487-1494, Sept 1972.

[27] T. G. Stockham, Jr., T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," Proc. IEEE, vol 63, pp678-692, April 1975.

[28] P. C. Baudelaire, "Digital Picture Processing and Psychophysics: A Study of Brightness Perception," Ph.D. Dissertation, Comput. Sci. Dep., Univ. Utah, Salt Lake City, 1972.

[29] H. Davis and S. R. Silverman, Hearing and Deafness. New York: Holt, Rinehart and Winston, 1970.

ACKNOWLEDGEMENTS

I wish to thank the people who have helped me in the course of the research leading to this dissertation. The late Professor J. W. Keuffel first interested me in research on speech and hearing because of his concern with the problems of the deaf. His support was essential in allowing me to undertake this research. Professor T. G. Stockham, Jr. supervised this dissertation. Much of the work described here is based on his original research in speech and image processing. Many others provided ideas, encouragement, and technical support. Particular thanks are due to Dick Warnock and Barden Smith, who kept the equipment running, to Mike Milochik, who printed all of my spectrograms, and to George Randall and Brent Baxter, who kept me honest - most of the time.