DIGITAL AUDIO PROCESSING TRAINING MANUAL

Third Edition

THE POWER TO HEAR – THE POWER TO CONVICT™

DIGITAL AUDIO PROCESSING TRAINING MANUAL

Third Edition

James Paul, Ph.D.

Donald Tunstall

DIGITAL AUDIO CORPORATION
4018 Patriot Drive, Ste. 300
Durham, NC 27703 U.S.A.

Telephone: 1-919-572-6767
Fax: 1-919-572-6786

www.dacaudio.com

Third Edition

August 17, 2004

Document Number 040817

Direct all inquiries to Digital Audio Corporation, 4018 Patriot Drive, Suite 300, Durham, NC 27703.

Privately published by Digital Audio Corporation

Printed in the United States of America

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. SOUND AND AUDIO

1.1 Sound Waves

Sound waves are *time-varying* pressure disturbances in a medium, usually air, that have energy in the human hearing frequency range. Disturbances at frequencies above and below this range are known as *ultrasonic* and *infrasonic* sounds, respectively.

Sound, or acoustic, waves in air are *longitudinal* waves that result from sympathetic vibration of air particles. The net effect is a wave radiating out from the sound source as compressions and rarefactions of the air. The vibrating particles do not move from their nominal positions but impart energy to adjacent particles, resulting in peaks and valleys in the air density.

Figure 1-1 illustrates acoustic wave compression (pressure increase) and *rarefaction* (pressure reduction) along with a plot of air pressure, *relative* to normal atmospheric pressure.

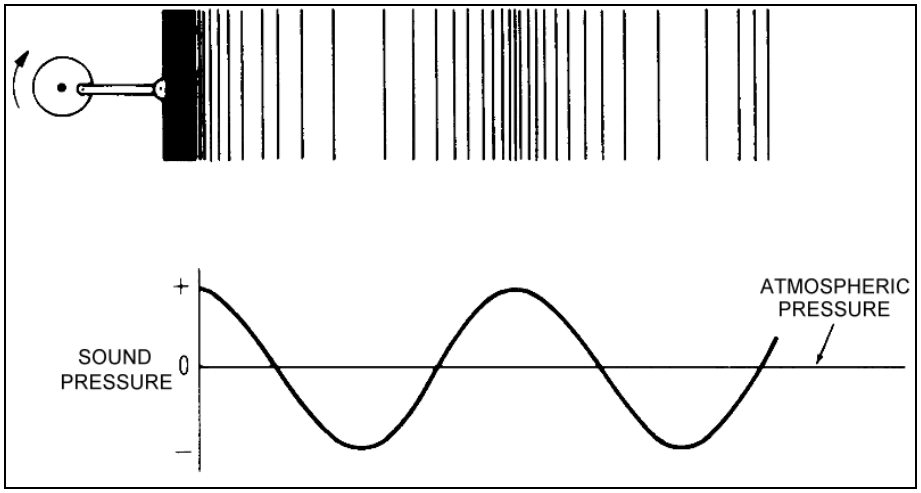


Figure 1-1: Illustration of Acoustic Wave Propagation and Pressure

The wavelength λ of a sound is determined by its frequency *F* and the speed of sound *α*, *i.e.,*

$$\lambda = \alpha / F$$

For example, the wavelength of a 500 Hz tone in air at 21 °C is

$$\lambda = 344 \text{ m/s} \div 500 \text{ Hz} = 0.68 \text{ meters or } 27 \text{ inches}$$

Sound travels through air at a speed of 1130 feet per second (at 21°C). Sound speeds in various media are given in the table below. Note that cooler air reduces the speed of sound and the wavelength becomes shorter. In addition, sound waves travel much faster through solids and liquids than through gases.

Table 1: Speed of Sound in Different Media

Medium	Meters/Second	Feet/Second
Air, 0°C	331	1087
Air, 21°C	344	1130
Water, fresh	1480	4860
Water, salt, 21°C, 3.5% salinity	1520	4990
Plexiglas	1800	5910
Wood, soft	3350	11000
Fir timber	3800	12500
Concrete	3400	11200
Mild steel	5050	16600
Aluminum	5150	16900
Glass	5200	17100
Gypsum board	6800	22310

1.1.1 Sound Pressure Measurements (dBA SPL)

Sound pressure level (SPL) is measured in *decibels,* or dB (see section 1.3, below), defined as

$$\text{dB SPL} = 20 \log \left(\frac{\text{sound pressure}}{0.00002 \text{ N/m}^2} \right)$$

The sound pressure is measured in Newtons per square meter. The level of 0.00002 N/m^2 (0 dB SPL) roughly corresponds to the threshold of hearing at 1000 Hz. Sound pressure meters are commonly available for carrying out these measurements. These meters use a *sound weighting curve* (Figure 1.) which gives more emphasis to certain frequencies. The A, B, and C curves approximate hearing sensitivities, which correspond to perceived equal loudness (see Figure 2-1).

The A curve is most universally used, though the B and C curves are often used at higher sound pressure levels. Measurements made using weighting are called "dBA sound pressure level" or "dBA SPL." Typical levels encountered by humans are given in Table 2.

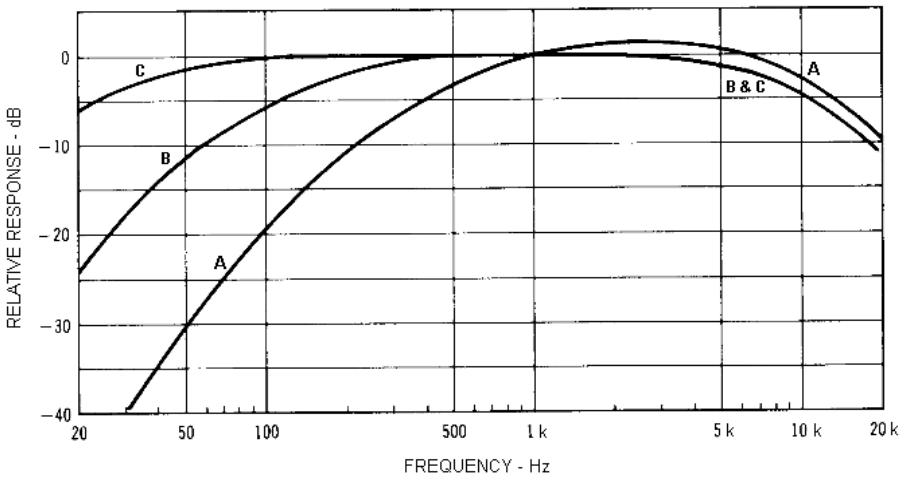


Figure 1.2: Sound Weighting Curves

Table 2: Typical Sound Pressure Levels

Level (dBA SPL)	Example
135	threshold of pain
125	jack hammer
115	car horn
95	subway train
75	street traffic
65	conversation
55	business office
45	living room
35	library reading room
25	bedroom at night
15	broadcast studio
5	threshold of hearing

1.2 Audio

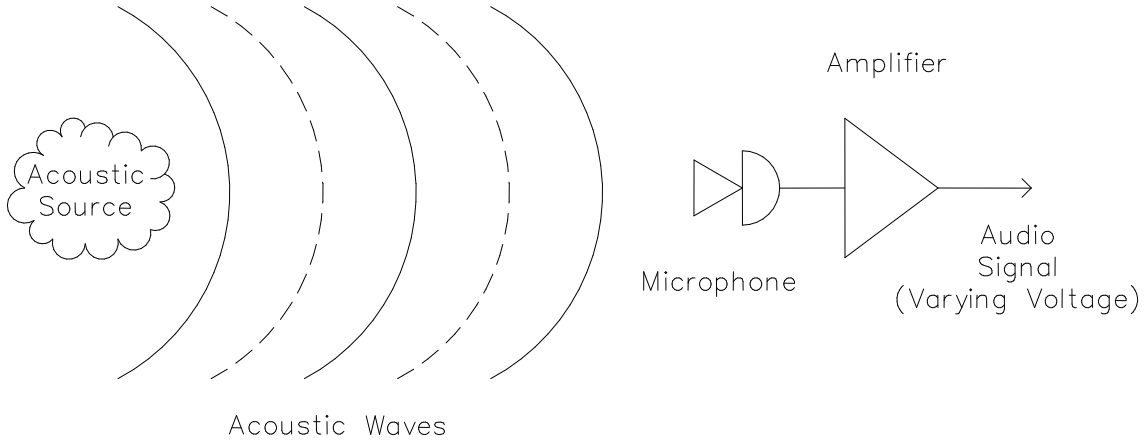


Figure 1.3: Sound Waves Converted to an Audio Signal

As illustrated in Figure 1.3, a microphone is a device that converts the time-varying mechanical energy of a sound wave into time-varying electrical energy known as *audio*. There are numerous technologies that can convert sound into electrical energy; chapter 7 provides details on the various types of microphone systems that are available.

Because the audio voltage produced by a microphone system is directly *analogous* to the sound wave, we refer to this as an *analog signal.* Audio signals produced by microphones are generally quite small, of the order of 1 millivolt (known as "mic level"), and so must be amplified to something on the order of 1 volt (or "line level") to be used effectively by audio instruments. We refer to this process as *preamplification.*

Audio levels are measured in either *volts* or *decibels* (dB). Since voice and other audio signals are dynamic in nature, *i.e.*, their levels fluctuate widely, measurements are usually made on the loudest segments of audio.

Audio instruments are *peak* sensitive. Their audio input gain controls are normally adjusted to accommodate the loudest sounds. Excessively loud audio will overload the instrument causing distortion. All digital audio instruments, such those produced by DAC, are especially sensitive to input overload, as analog-to-digital converters produce sharp *clipping* distortion on overload.

Audio signals are actually alternating voltages and have to be measured with alternating current (AC) meters. Alternating voltages swing both positive and negative. Figure 1.4 illustrates a *time waveform* of a pure tone (sine wave). As time progresses (moves to the right along the time axis), the waveform increases to a positive peak voltage, falls through zero voltage, and falls to a peak negative voltage. Direct current (DC) voltages do not fluctuate with time.

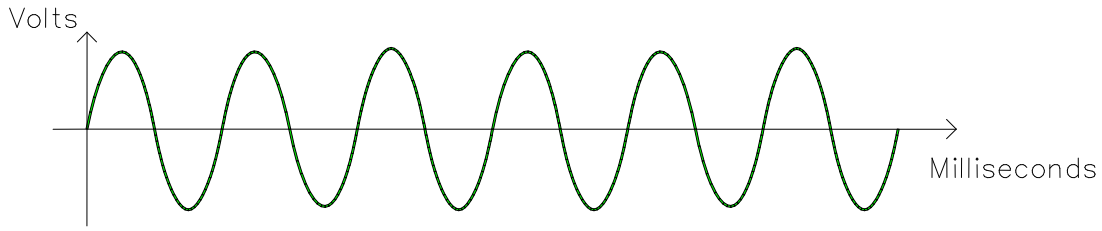


Figure 1.4: Audio Waveform of a Tone (Sine Wave)

AC voltages such as audio are generally expressed in terms of *root-mean-square* (RMS) voltage, which, unlike the peak voltage that actually gets measured, compares directly with DC voltage of the same value.

For example, an AC voltage of $12V_{RMS}$ imparts the same power to a given load as does 12VDC. For any sine wave, which is the simplest type of AC voltage:

$$V_{RMS} = 0.707 \times \text{peak voltage}$$

So a $12V_{RMS}$ actually has a peak voltage of 17V.

Example:

Household voltage is measured to be 120 V_{RMS}.
What is the peak voltage of this sinusoid?

$$V_{RMS} = 0.707\ V_{peak}$$

$$V_{peak} = V_{RMS} / 0.707$$

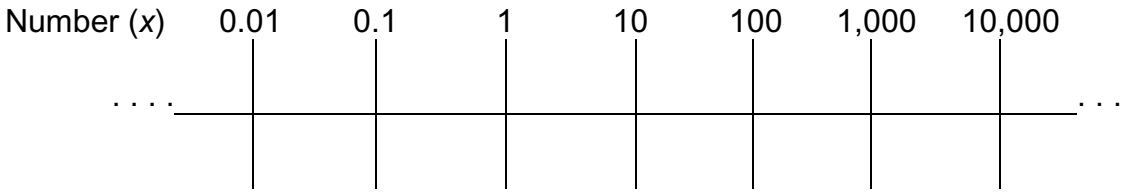
$$V_{peak} = 120\ V / 0.707 = 170\ V$$

The AC outlet actually produces peak voltages up to 170 volts.

Again, AC voltage is usually expressed simply as *volts,* with the understanding that it is RMS, not peak.

1.3 Decibels

The human ear hears sound levels on a compressed scale. Sounds having twice as much energy are perceived as less than twice as loud. The ear's sensitivity very closely resembles a *logarithmic* scale. Figure 1. illustrates the compression effect of the log function.



log *x*	-2	-1	0	1	2	3	4
10 log *x*	-20	-10	0	10	20	30	40
20 log *x*	-40	-20	0	20	40	60	80

Figure 1.5: Logarithmic Compression Effect

Note that the log of a number is its tens exponent, *e.g.*,

log (1000) = log (10^3) = 3, and

log (0.001) = log (10^{-3}) = -3

The log of a number is usually derived by using a scientific calculator or computer, as it involves a complex sequence of computations.

Example: Find the logarithms of 150, 0.065, 0, and -1.0

The log of 150, from Figure 1., lies between 2 and 3.
Using a calculator, log (150) = 2.176.

Similarly, log(0.065) = -1.187.

The log of 0 is actually -∞ and will produce an error on the calculator. The log of any negative number is *not* defined and is, therefore, meaningless. Logarithms are taken only of positive numbers. For mathematical purists, the log of a negative number actually yields a complex number with real and imaginary components, but here we are concerned only with real numbers.

Decibels (dB) measure voltage, amplification, and sound levels. In audio, voltage measurements are the most popular form of dB.

1.3.1 Voltage Decibels (dBm and dBV)

Audio levels are most often measured in decibels (dB), defined as follows:

$$dB = 20 \log \left(\frac{volts}{V_{ref}} \right)$$

Note that a logarithm is taken of a *reference voltage*, V_{ref}, which specifies the type of decibel being used. The most popular audio unit is dBm and has a reference voltage of 0.775 volts. The "m" in dBm refers to 1 milliwatt into a 600 ohm load (0.775 volts from Ohm's law) and was originally developed by the telephone industry.

Because modern audio equipment may utilize loads other than 600 ohms, the term dBu is often used interchangeably with dBm:

$$dBm = dBu = 20 \log \left(\frac{volts}{0.775} \right)$$

Another less popular, though arguably more logical, audio measurement unit is dBV; it utilizes a reference voltage of 1.0 volt.

$$dBV = 20 \log \left(\frac{volts}{1.0} \right), \text{ or simply}$$

$$dBV = 20 \log (volts).$$

dBV is popular for instrumentation measurements, and is commonly used in spectrum analyzers.

Consider the following example:

Example: Express 3.9 volts in both dBm and dBV.

$$\text{dBm} = 20 \times \log\left(\frac{3.9}{0.775}\right) = 20 \times \log(5.03)$$

$$= 14.04 \text{ dBm}$$

$$\text{dBV} = 20 \times \log\left(\frac{3.9}{1}\right)$$

$$= 11.82 \text{ dBV}$$

From this example we learn that an audio level expressed in dBm is always 2.2 dB greater than the dBV version of the same voltage, *i.e.*,

$$\text{dBm} = \text{dBV} + 2.2$$

Converting dBV and dBm back to V_{RMS} is slightly more complicated.

The following equations allow converting dB back to volts:

$$\text{volts} = 0.775 \times 10^{\text{dBm}/20}$$

$$\text{volts} = 10^{\text{dBV}/20}$$

Example: Convert +15 dBV to AC voltage.
Convert –20 dBm to AC voltage.

$$\text{volts} = 10^{15/20}$$

$$\text{volts} = 10^{0.75}$$

$$= 5.62 \text{ volts RMS}$$

$$\text{volts} = 0.775 \times 10^{-20/20}$$

$$= 0.775 \times 10^{-1}$$

$$= 0.076 \text{ volts RMS}$$

1.3.2 Volume Units (VU)

Tape recorder meters often express audio levels in volume units (VU). A VU meter is an AC voltmeter calibrated on a dB scale with a ballistically dampened movement. The dampening smoothes out peaks caused by abrupt peaks in audio signals. Without this dampening, the meter would vibrate rapidly on audio signals and not be useful in making measurements.

The meter is usually calibrated such that 0 VU represents the loudest signal acceptable to the recording tape. For sine-wave measurements, a change of 1 VU is equivalent to a change of 1 dB in amplitude.

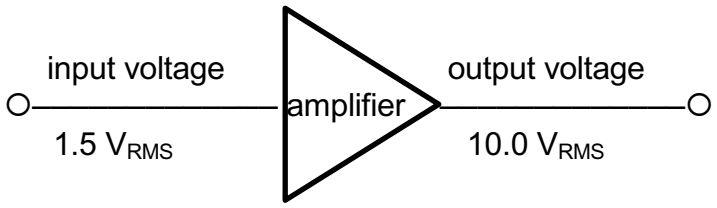
1.3.3 Amplification Decibels (dB)

The gain of voltage and power amplifiers is also measured in dB. A voltage amplifier is commonly used as a microphone or line amplifier where the goal is to merely increase (or decrease) the voltage level of the audio. A power amplifier is used to drive a loudspeaker.

The dB gain of a *voltage* amplifier is expressed as

$$\text{Gain} = 20 \log \left(\frac{\text{output voltage}}{\text{input voltage}} \right)$$

Example: What is the dB gain of the following amplifier?



$$\text{Gain} = 20 \log \left(\frac{10.0}{1.5} \right)$$

= 16.5 dB of gain

When amplifiers are cascaded (seriesed), as in Figure 1-2 below, the dB gains are *added*.

Example: What is the dB gain of the circuit shown?
What is the output audio level in dBm, dBV?

Figure 1-2: Seriesed Voltage Amplifiers

$$A_1 \text{ gain: } 20 \log \left(\frac{0.1}{0.001} \right) = 40 \text{ dB}$$

$$A_2 \text{ gain: } 20 \log \left(\frac{5}{0.1} \right) = 34 \text{ dB}$$

Amplifier gain: Gain = 40 dB + 34 dB
= 74 dB

The dB gain of a *power* amplifier is expressed as

$$\text{dB power} = 10 \log \left(\frac{\text{Power out}}{\text{Power in}} \right)$$

The following example illustrates the dB gain of a *power* amplifier.

Example:

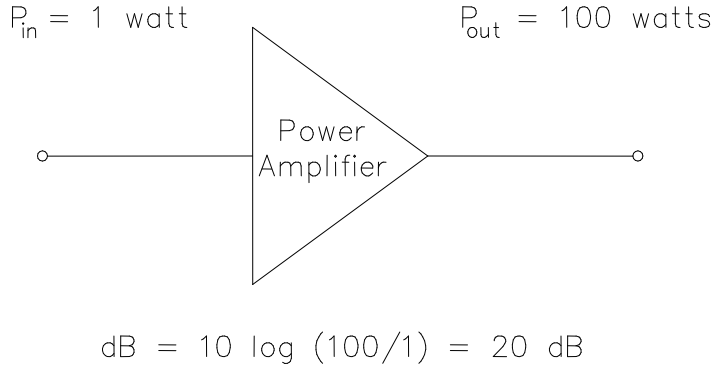


Figure 1-3: Power Amplifier

1.4 Audio Line Levels

An audio instrument such as a DAC filter or a tape recorder usually has audio input and output connectors that produce *line level* audio. These connectors are often referred to as line inputs and

line outputs. Such line level connections are intended for interconnection of instruments and are not intended for driving headphones or a loudspeaker. Audio line levels are nominally around 1 volt (2.2 dBm or 0 dBV) but may range from 0.5 volts to 5 volts, depending upon the instrument.

1.5 Three-dB Bandwidth

The bandwidth of an audio instrument is often referred to as its "3 dB bandwidth." The 3 dB point, shown below,

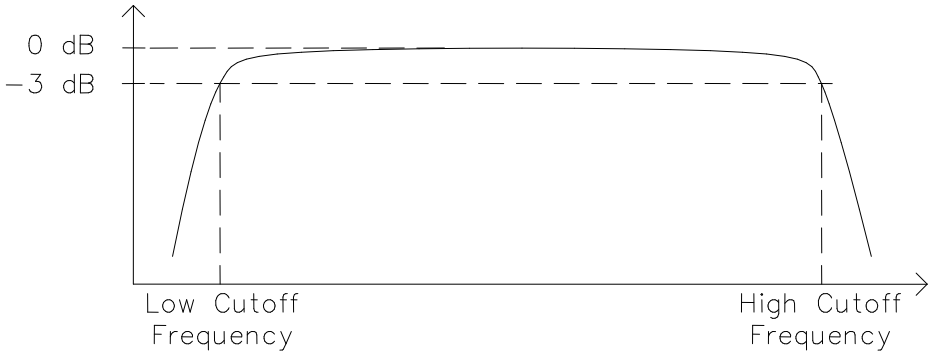


Figure 1-4: Three-dB Bandwidth

is the High Cutoff Frequency minus the Low Cutoff Frequency. Three decibels is of particular interest since it represents the point at which the midband power level drops to one-half, *e.g.*,

$$10 \log \left(\frac{0.5 \text{ midband power}}{\text{midband power}} \right) = 10 \log (0.5) = -3.01 \text{ dB}$$

1.6 Units

Audio processing describes voltages and frequencies in units with prefixes as follows:

Giga (G)	10^9 or 1,000,000,000
Mega (M):	10^6 or 1,000,000
Kilo (k):	10^3 or 1,000
milli (m):	10^{-3} or 0.001 or 1/1000
micro (μ):	10^{-6} or 0.000 001 or 1/1,000,000

Examples:

6.3 μV (microvolts) = 0.000 006 3 volts

17.21 kHz (kilohertz) = 17,210 Hz

1.7 dB Computations without a Calculator

Audio voltage (and amplifier gains) may be easily converted relatively accurately without a calculator. The following five rules can be memorized to carry this out.

1. 0 dBm is 0.775 volts
   0 dBV is 1.0 volts

2. An increase of 10× in voltage corresponds to adding 20 dB.

   Examples:

   100 volts = 1 volt × 10 × 10
   = 0 dBV + 20 dBV + 20 dBV
   = 40 dBV

   7.75 volts = 0.775 volts × 10
   = 0 dBm + 20 dBm
   = 20 dBm

3. A decrease in voltage by 1/10 corresponds to subtracting 20 dB.

   Example:

   775μV = 0.775 × 1/10 × 1/10 × 1/10
   = 0 dBm – 20 dBm – 20 dBm – 20 dBm
   = –60 dBm

4. Doubling (or halving) the voltage adds (subtracts) 6 dB.

   Examples:

   4 volts = 1 volt × 2 × 2
   = 0 dBV + 6 dBV + 6 dBV
   = 12 dBV

   388 mV = 0.775 × 1/2
   = 0 dBm – 6 dBm
   = –6 dBm

5. To convert from dBm to dBV, subtract 2.2 dB.
   To convert from dBV to dBm, add 2.2 dB.

   Example from above:

   4 volts = 12 dBV= 14.2 dBm
   388 mV = –6 dBm = –8.2 dBV

EXERCISES

1. A loudspeaker in an open area is producing 0.1 watt per square foot at 10 feet distance. What is the sound intensity at 20 feet? At 40 feet?

2. What is the RMS voltage of a 1000 Hz tone having a peak voltage of 2.83 volts? What is the RMS voltage if the frequency is changed to 2000 Hz?

3. A microphone is rated at an output level of –60 dBm. What is the equivalent RMS voltage output?
   How much amplification, in dB, is required to boost this level to +6 dBm?

4. A 1 watt power amplifier produces a sound pressure level of 65 dBA SPL. How much power is required for 85 dBA SPL?

5. A 1 kHz tone registers 65 dBA SPL. The frequency is changed to 100 Hz, and the sound intensity is not changed. What will the meter register in dBA, dBB (B scale), and dBC (C scale)?

6. An amplifier outputs 3 V_{rms} when its input audio measures 30 mV_{rms}. What is its gain in dB? If the input level is increased to 50 mV, what is the output level in V_{rms}, dBm, and dBV?

7. Without using a calculator, convert the following:
   a. 1 V_{rms} to dBV and dBm?
   b. +20 dBm to V_{rms}?
   c. –40 dBV to dBm and V_{rms}?
   d. 0.5 V_{rms} to dBm?

2. AUDIO FREQUENCY MEASUREMENTS

2.1 Audio Frequency Range

The human ear, under ideal conditions, can sense acoustic energy over the approximate range of 20 to 20,000 hertz. A hertz, abbreviated Hz, is the measurement of frequency and replaces what was once *cycles per second* (cps). Kilohertz (kHz), a unit of 1000 Hz, often is used in higher frequency measurements.

Although the human ear perceives a wide range of frequencies, it does not respond the same at all frequencies. In Figure 2-1, the Fletcher-Munson curves show contours of equal loudness level in terms of dB SPL.

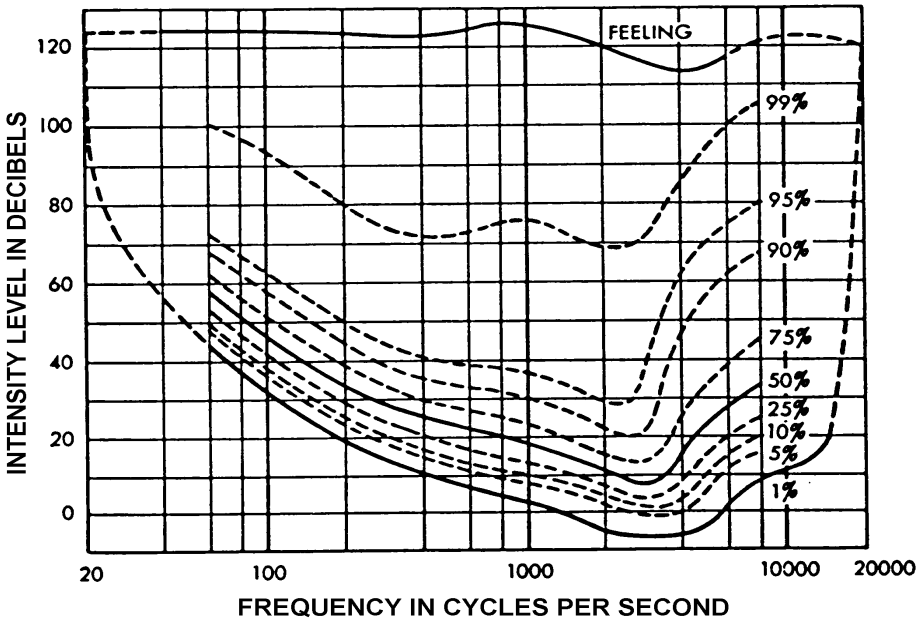


Figure 2-1: Fletcher-Munson Equal Loudness Curves

Note that at the ear's peak *dynamic range*, the separation between the faintest sounds and strongest sounds (before pain), occurs at approximately 3000 Hz and is in excess of 120 dB. This corresponds to an amplitude variation of 1,000,000:1! At high and low frequencies this *dynamic range* becomes much smaller.

Forensic audio processing usually relies on a smaller range of frequencies, typically 200 Hz to 5000 Hz. The bulk of voice energy is concentrated in this range. The 200 to 5000 Hz frequency

range, or *voice spectrum*, usually contains all acoustic information necessary for the forensic technician. Audio energy below 200 Hz carries little speech information and often contains room resonances and background rumble. Energy above 5000 Hz often contains disproportionate amounts of high frequency noises, *i.e.*, a poor signal-to-noise ratio. Although the pitch of a male speaker's voice can go as low as 50 Hz, the presence of harmonics lets us perceive the pitch even without directly capturing it.

2.2 Frequency vs. Time

Audio signals may be displayed in either the *time domain* or the *frequency domain*. Time domain displays are the actual waveform, *i.e.*, the voltage fluctuations displayed with time. An oscilloscope displays in the time domain. The frequency domain displays the energy in the audio at different frequencies. Spectral analysis is the process whereby the energy content of a signal at different frequencies is measured. Pure tones are the only type of audio that has all of the energy concentrated at a single frequency. Consider the tone's waveform and its *frequency spectrum* in the figure below.

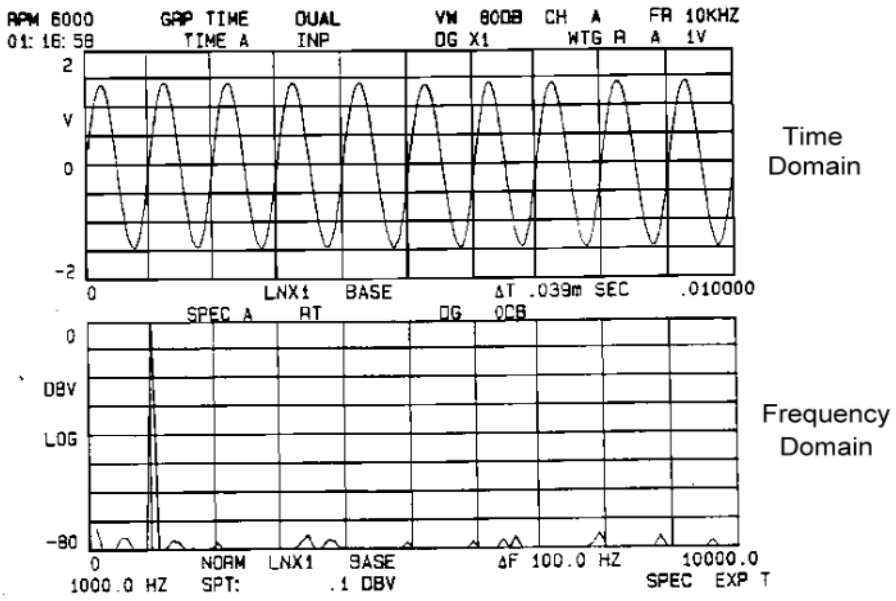


Figure 2-2: Time and Frequency Representations of a Tone

The tone illustrated in Figure 2-2 has a repeat interval, called a *period*, of 0.001 seconds (1 millisecond). Since the period repeats itself 1000 times in one second, its frequency is 1000 Hz.

The bottom display of Figure 2-2 gives the frequency domain (spectrum) representation of that waveform. Note that the energy is concentrated at 1000 Hz, and its level is 0 dBV. From Section 1.7, 0 dBV corresponds to 1 volt (RMS). Section 1.2 says that a 1 volt (RMS) sinewave has a peak value of 1.414 volts. Confirm this peak voltage in the top half of Figure 2-2.

The relationship between the period of a tone and its frequency is

period (sec) = 1 / frequency (Hz)

frequency (Hz) = 1 / period (sec)

Example: Find the period of 60 Hz AC power

period = 1 / 60 = 0.0167 sec = 16.7 msec

Find the frequency of a tone whose waveform repeats itself every 5 msec.

frequency = 1 / 0.005 = 200 Hz

What is the frequency of the earth's rotation about the sun?

frequency = 1 / (365 × 24 × 60 × 60)
= 0.000000032 Hz

The tone waveform shown in the figure is called a *sine wave* because it is described by the trigonometric function

$$v(t) = V_{peak} \sin (360 \cdot f \cdot t),$$

where $v(t)$ is the voltage at a time t seconds from the start of the time scale, f is the frequency of the tone, and V_{peak} is the maximum (peak) voltage of the sine wave.

Example:

A tone has an RMS voltage of 1.414 volts and a frequency of 100 Hz. What is the equation of the voltage waveform? What is its value at 5 msec, at 12 msec?

From Chapter 1,

$V_{peak} = V_{RMS} / 0.707$

Equation: $v(t) = 2.0 \sin (36000\ t)$

Voltage at 5 msec: $v(0.005) = 0$ volts

Voltage at 12 msec: $v(0.012\text{sec}) = 1.9$ volts

Often the terms tone and sine wave are used interchangeably.

Audio, such as voice, noise, and music, has its energy spread over a continuous range of frequencies. Unless tones are present, the power spectrum appears as a distribution of audio energy spread across frequencies rather than discrete lines. Figure 2-3 illustrates a typical voice audio spectrum averaged over a long period of time.

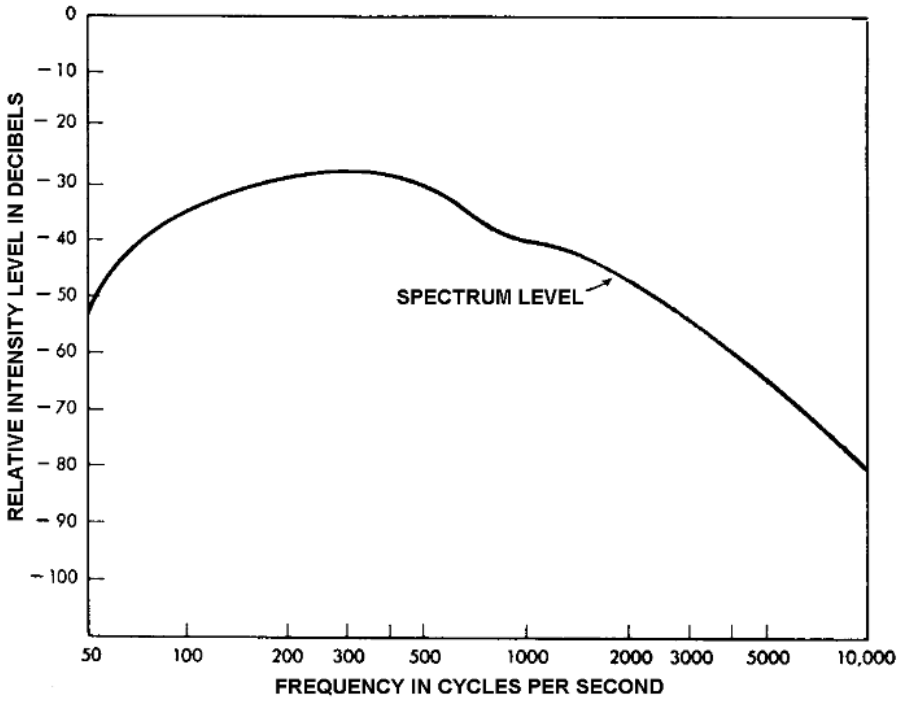


Figure 2-3: Typical Averaged Voice Spectrum

2.3 Spectrum Analysis

A *power spectrum analyzer* is an instrument used to measure the power spectrum of a signal. Such instruments are special digital signal processors which enable the user to display visually the power spectrum on a CRT. These instruments measure the power spectrum of a waveform,

such as an audio signal, using the Fourier transform, a special mathematical equation converting the *time domain* to the *frequency domain*. The actual implementation procedure (or algorithm) used is the *fast Fourier transform* (FFT). As result, such instruments are referred to as FFT analyzers.

The CRT display for a Stanford Research Systems SR760 analyzer is given in Figure 2-4. Power spectrum analyzers are general purpose instruments that are used by a variety of disciplines including machinery analysis, vibration, acoustics, and speech analysis. FFT analyzers have a plethora of controls, but the most important are:

- Bandwidth – the frequency range displayed
- Resolution – the number of frequency points displayed across the bandwidth
- Dynamic range – the weakest to the strongest energy displayed
- Averaging – the smoothing used to moderate time fluctuations in the displayed spectrum

Additionally, controls permit adjusting the input signal levels to the electronics and the peak signal level displayed on the CRT screen.

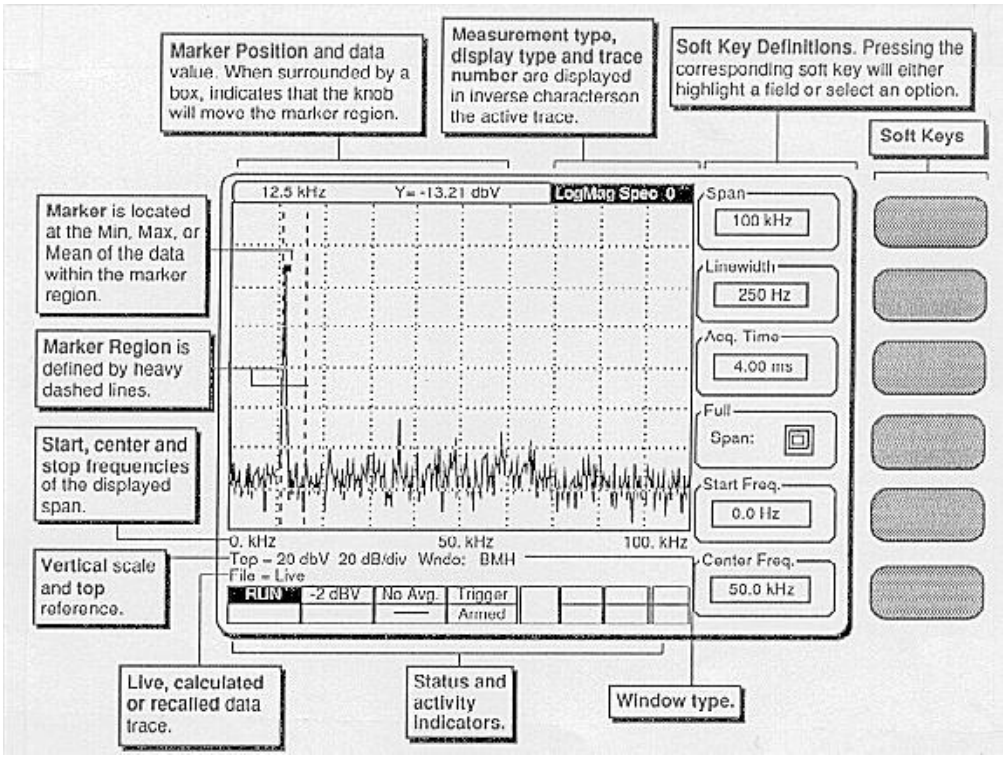


Figure 2-4: SR760 Screen Display

FFT power spectrum analysis is equivalent to measuring and smoothing the energy in very narrow frequency bands across the audio spectrum. The energy in each band is then plotted on a CRT as a sequence of dots. The dots are usually connected with lines so as to give the impression of a continuous curve. Figure 2-5 *functionally* illustrates this process.

In Figure 2-5 the audio signal is partitioned into 400 equal-width, contiguous frequency segments. Bandpass filters (BPFs) are used. The outputs of these 400 BPFs are each rectified to permit power measurement in each band and are then smoothed with individual averages. If no smoothing were to take place, the *real-time* spectrum would jump very rapidly with the dynamic audio. Measurements would be difficult to make.

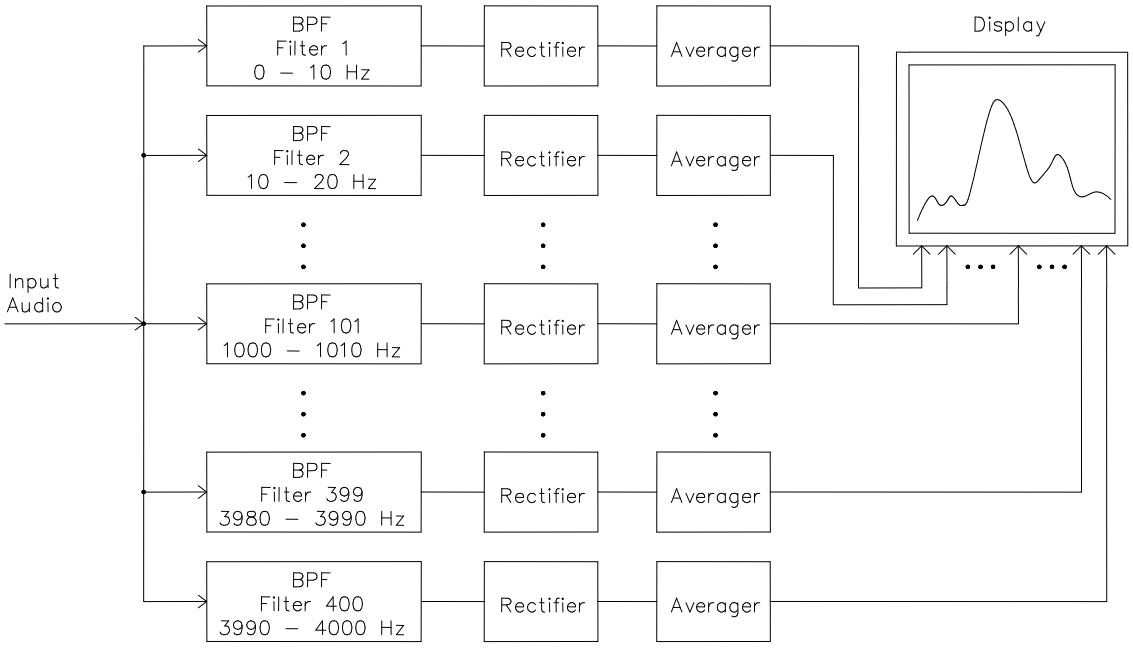


Figure 2-5: 400-Band Power Spectrum Analyzer

The smoothing used is normally *exponential averaging*, which is one which weights the current signal power the most and recent signal power less. The older the signal, the less it influences the current display.

Smoothing is normally used for speech analysis. Since the voice spectrum is nonstationary (changes continuously from one sound to the next), an average spectrum is preferred. Deviations from the voice's smoothed spectrum (Figure 2-3) indicate added noises and acoustic modifications.

Note that (Figure 2-5) the 101st spectral *line* results from a bandpass filter isolating the signals energy between 1000 and 1010 Hz. If a 1003 Hz tone (all energy concentrated at a 1003 Hz frequency) were present it would register on the CRT display at 1005 Hz (the center of that filter). Actually, any tone from 1000 to 1010 Hz would indicate 1005 Hz. The bandpass filters

are not ideal, *i.e.*, they do not partition the spectrum into perfectly isolated 10 Hz-wide bands. The tone at 1003 Hz will also *leak* into the next lower band centered at 995 Hz (Filter 100).

The actual operation of the FFT power spectrum analyzer is illustrated in Figure 2-6 and is slightly different from the illustration above. Bandpass filters are not used. Instead, the audio signal is partitioned into short segments. Each segment is *windowed* by a weighting curve which emphasizes the center of the segment and feathers either end of that segment to zero.

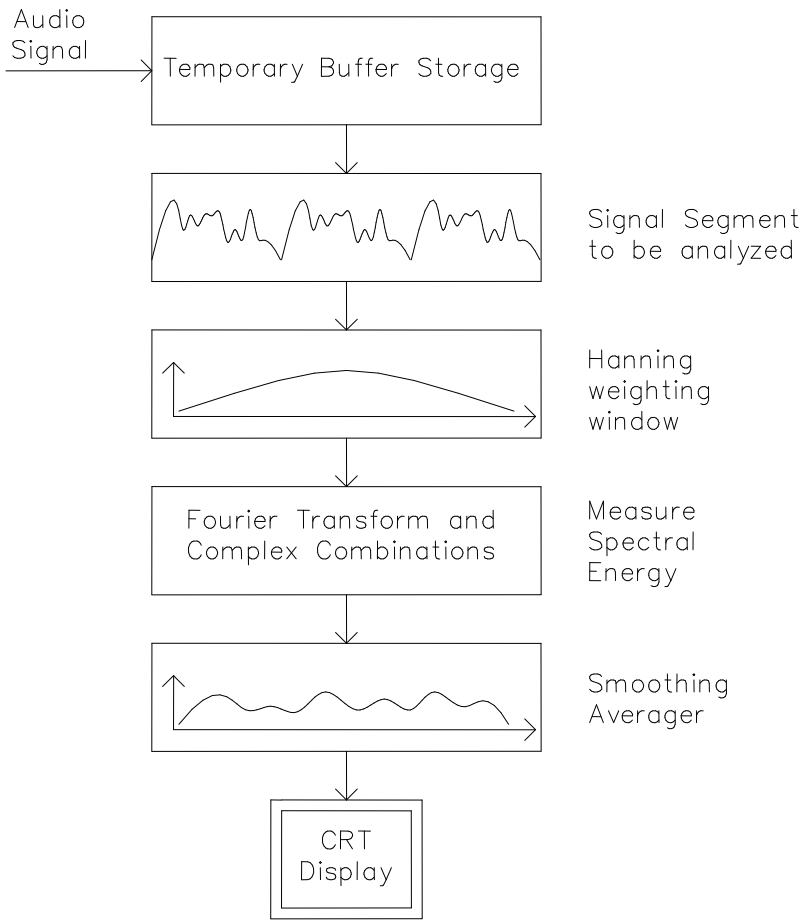


Figure 2-6: FFT Power Spectrum Analyzer

The windowed segment is processed by a Fourier transform algorithm in a computer. This processing results in spectral energy outputs which are equivalent to the bandpass filter outputs of Figure 2-5. The actual Fourier transform operation is beyond the scope of this material, and also requires the combination of complex numbers. The big advantage of using an FFT is that it requires fewer mathematical operations than the bandpass filter approach and thus is faster.

The weighting window used on each segment specifies the shape of the equivalent bandpass filters. Different windows have different properties, but all attempt to reduce energy leakage between spectral lines and maximize accuracy. Two excellent windows for audio are the

Hanning and Blackman windows. Never use a flat-top, or rectangular, window for audio analysis due to its poor leakage characteristics.

A *short-term power spectrum* is produced by the Fourier transform for each segment. Audio is a dynamic process, and its characteristics change rapidly with time; therefore, the sequence of short-term spectra also changes rapidly. Information is difficult to extract from the display of dynamically changing spectra. As a result, smoothed spectra, produced by averaging the short-term spectra, are displayed. The averaging process, known as *exponential averaging*, gives greater weight to current short-term spectra and less to older ones. The effective number of averages is controlled by the user.

The FFT spectrum analyzer's bandwidth is usually set to the bandwidth of the signal being analyzed. A telephone recording, for example, has little energy above 3 kHz. The analyzer would be set to the next available bandwidth greater than 3 kHz. Commercial analyzers have bandwidths adjustable in several steps from approximately 100 Hz to 100 kHz.

A valid reason for not over specifying the analyzer's bandwidth is to provide greatest resolution. Typically, analyzers give 400 or 800 *lines* of resolution (display points). Each bandpass filter in Figure 2-5 produces a single line of resolution. A 400-line analyzer with 4 kHz bandwidth gives an energy measurement every

$$4000 \text{ Hz} / 400 = 10 \text{ Hz}.$$

Each displayed point on the CRT represents the energy in a 10 Hz slice of the power spectrum. Normally, finer resolution displays give better insight into noise characteristics and allow more precise specification of tone frequencies.

Popular FFT spectrum analyzers used in forensic analysis are the Scientific Atlanta SD390 and the Stanford Research Systems SR760 and SR770.

2.4 Octave Analysis

Octave-based spectrum analyzers are also available but are not normally recommended for the audio analysis addressed here. Such systems typically partition the audio spectrum into 25 bands, with bandwidths increasing with center frequency. Such instruments are most suitable for analysis of sound reinforcement systems and have poor frequency resolution, especially at higher frequencies.

Often the term "octave" is used in conjunction with frequency analysis. An octave up in frequency is double that frequency; an octave down is half that frequency. For example,

Frequency:	1000 Hz
One octave up:	2000 Hz
Two octaves up:	4000 Hz

Frequency:	1000 Hz
One octave down:	500 Hz
Two octaves down:	250 Hz

Analog filter rolloffs are often expressed as "dB per octave." A lowpass filter with a rolloff of 24 dB per octave and a cutoff frequency of 1000 Hz has an attenuation of 24 dB at 2000 Hz and 48 dB at 4000 Hz.

EXERCISES

1. What is the period of a 1 kHz tone?
   What is the frequency of a tone with a period of 2 µsec?

2. The SR770 spectrum analyzer has 400 lines of resolution.
   What is its frequency resolution at 3.125 kHz bandwidth? At 6.25 kHz?
   Which bandwidth would be most suitable for analyzing a body transmitter having 4.0 kHz audio bandwidth?

3. What frequency is one octave up from 1000 Hz?
   What frequency is one octave down from 1000 Hz?

3. BASIC ELECTRICAL CIRCUITS

3.1 Volts, Amps, and Watts

An electrical source, such as a battery, has an electric potential known as its *voltage* (*V*). The volt (V) is the standard unit of electric potential. A *current* (*I*) exists when electrons flow through a circuit. The ampere (A), also called the amp, is the standard unit of measure for electric current. Voltage is analogous to the pressure of a stream of water in pounds per square inch, whereas current resembles the rate of flow in gallons per hour.

Consider the following *schematic* diagram of a flashlight:

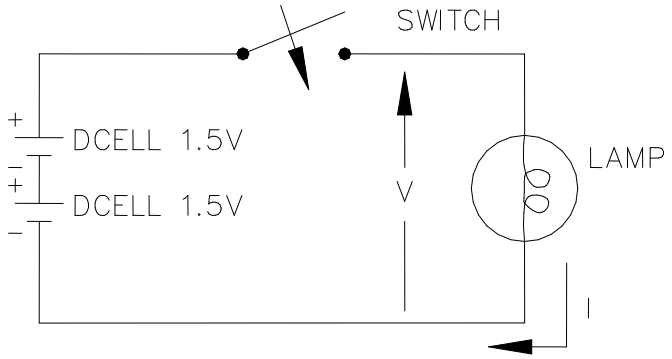


Figure 3-1: Schematic Diagram of a Flashlight

Two D cells, each with a voltage of 1.5 volts, are placed in series to produce a 3.0 volt potential. When the switch is closed to complete the circuit, a current *I* flows out of the positive (+) terminal through the lamp and back into the negative (–) terminal of the *battery*, *i.e.*, the two seriesed cells.

Power (*P*) is the rate at which electric energy is expended. The watt (W) is the standard unit of power. In direct current (DC), a circuit's power is the product of voltage and current. If a current of 0.1 amps (100 milliamps) flows through the lamp in the flashlight in Figure 3-1, the bulb dissipates

$$P = V \cdot I$$

$$0.3 \text{ watts} = 3 \text{ volts} \cdot 0.1 \text{ amps}$$

The 300 milliwatts of power are dissipated in both light and heat.

If the battery has a *capacity* of 2 amp-hours, then the lamp would light for

$$20 \text{ hours} = \frac{2 \text{ amp-hours}}{0.1 \text{ amps}}$$

3.2 Resistance

Electric resistance (R) impedes the flow of current and results in the dissipation of power as heat. The ohm (Ω) is the standard unit of resistance. The flashlight's lamp has electric resistance, which specifies the ratio of voltage across the lamp compared to the current flowing through the bulb, *i.e.*,

$$R = \frac{V}{I} \qquad \text{Ohm's law}$$

Ohm's law may be rewritten by rearranging terms as

$$V = I \cdot R \text{ and } I = V / R$$

The resistance of the bulb in the previous example in Figure 3-1 is

$$R = \frac{3 \text{ volts}}{0.1 \text{ amps}} = 30 \text{ ohms}$$

As a result, if one of the D cells were removed and replaced by a jumper, the bulb would receive only 1.5 volts and its current would be 0.05 amps. Why?

Power can be related to resistance:

$$P = V \cdot I, \text{ and } V = R \cdot I.$$

Therefore,

$$P = I \cdot I \cdot R = I^2 \cdot R.$$

Alternatively,

$$P = V \cdot I, \text{ and } I = V / R.$$

Therefore,

$$P = V \cdot V / R = V^2 / R.$$

Example:

> Power from a generator needs to be transmitted to a location 1 mile away. 20 amps of current will pass over the power line, which has a resistance of 10 ohms. How much power will be lost in the power line?

$$P = I^2 \cdot R = (20\ \text{A})^2 \cdot 10\ \Omega = 2000\ \text{watts}$$

Resistances may be placed in parallel or series. See Figure 3-2. Their parallel and series equivalent resistances are given.

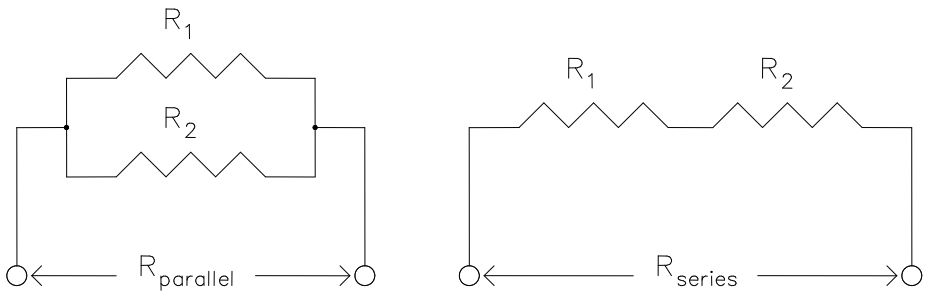


Figure 3-2: Parallel and Series Resistance Combinations

> Example: What is the equivalent resistance of a 10 ohm and 20 ohm resistor in series? In parallel?

Series: $R_{series} = R_1 + R_2 = 10\ \Omega + 20\ \Omega = 30$ ohms

Parallel: $$\frac{1}{R_{parallel}} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{1}{10} + \frac{1}{20} = 0.15$$

$$R_{parallel} = \frac{1}{0.15} = 6.7 \text{ ohms}$$

3.3 AC Circuits

Alternating current (AC) voltage, such as 120 VAC available at electrical outlets, is a voltage which varies in a sinusoidal fashion at a rate of 60 Hz in the U.S. or 50 Hz in most overseas countries. The equation for this voltage as a function of time is

$$v(t) = 170\ \text{V} \cdot \sin(3600\ t)$$

Rather than describe AC voltage with this awkward equation, the voltage is measured in terms of an equivalent DC voltage which would produce the same power in a resistor. As a result, a 120 V_{RMS} AC voltage (the equation above) would dissipate the same power in a resistor as a 120

VDC voltage. *Root-mean-squared* (RMS) is a mathematical expression for AC voltages equating their effect to DC voltages. Section 1.2 included further information on RMS voltages.

The AC equivalent to resistance is called *impedance* (*Z*). Actually, impedance consists of resistance and reactance. Reactance is produced by inductors and capacitors and is treated mathematically using complex arithmetic. Technically, impedance has both a magnitude and a phase angle. Discussion of this subject is beyond the scope of this document. As a simplification, impedance shall be viewed as resistance.

The equivalent AC expression for Ohm's law is

$$Z = \frac{V}{I} \ ,$$

where *Z* is impedance. Again, this expression may be rewritten as

$$V = I \cdot Z \quad \text{and} \quad I = V / Z.$$

Impedance is of particular interest here since *audio voltages are AC*. If the impedance Z is large, then the current is small. Many audio instruments have high *input impedances* to avoid unnecessary current *loading* of the audio source ($I = V / Z$).

Some audio instruments are configured for 600 ohm impedance. This impedance, which is a carryover from telephone technology, is most suitable when connecting audio instruments via a long cable or in the presence of electrical interference. The preferred impedances for forensic instruments are *low output* and *high input* impedances.

3.4 Ground Loops

Ground loops often occur when two or more AC powered audio devices are connected. A recorder connected to an amplifier will often introduce undesirable AC hum. This AC hum results when a small AC signal is introduced into an amplifier's input. The amplifier will increase the level of the AC hum along with the audio.

The AC signal may be inadvertently obtained from small differences (on the level of millivolts) in the AC ground potential at each instrument. Single-ended audio connections (center conductor plus shield) are particularly susceptible. Figure 3-3 illustrates the methods by which ground loops are produced. The most common causes are due to inductive coupling in a closed loop. In the figure, the changing magnetic field, usually stray AC magnetism from a nearby transformer or conductor, couples into the closed circuit formed by the AC safety ground and audio cable shield. A small 60 Hz voltage appears at the right amplifier. Hum is thus added to the audio.

Capacitive coupling and small voltage drops (due to equipment ground leakage currents) are also capable of introducing hum in audio circuits. Hum levels as low as 40 dB below voice levels are audible during silence periods. High-gain amplifiers, such as microphone and transformer amplifiers, are particularly susceptible to ground loops, as the hum components are also amplified.

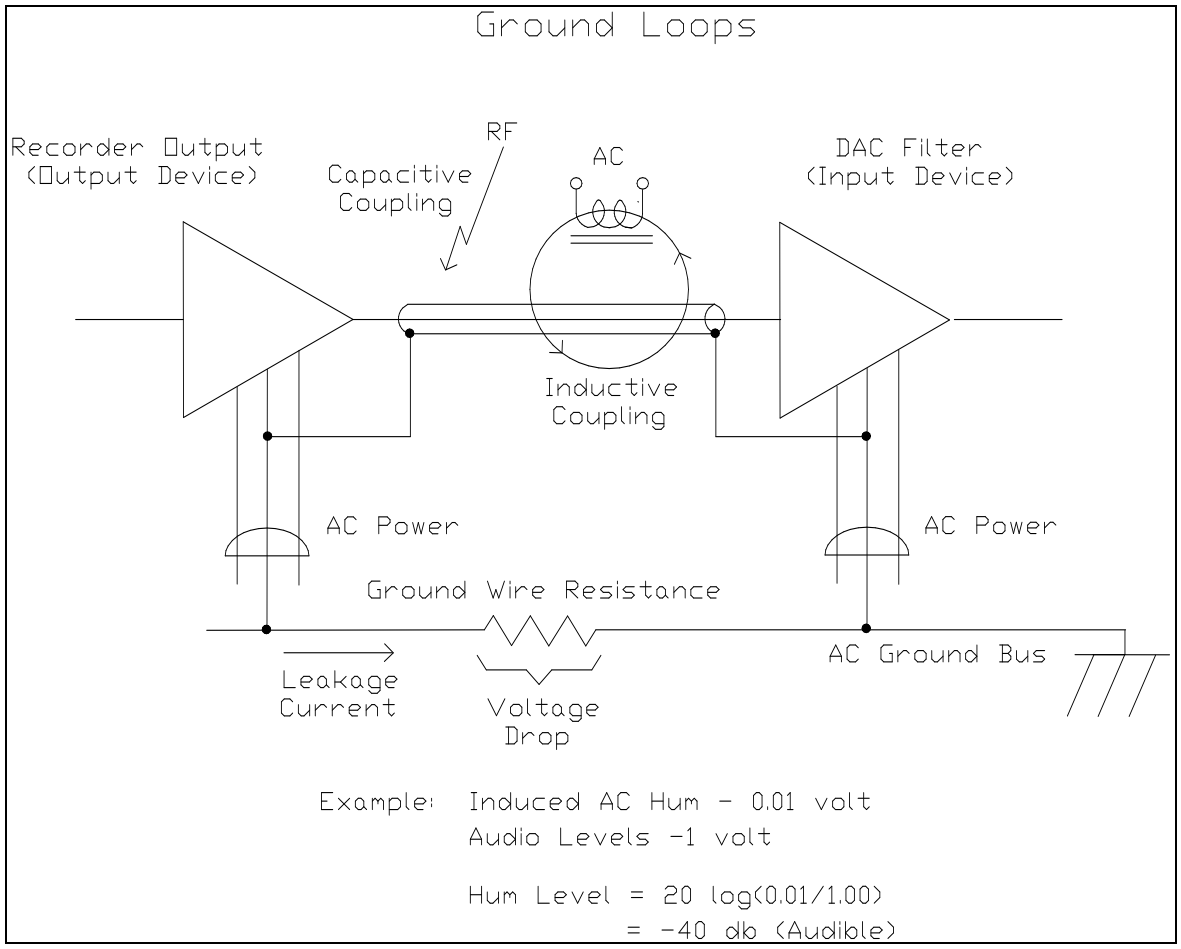


Figure 3-3: Ground Loop Sources

Fixing ground loops is something of a black art. The following suggestions may help.

1. Separate AC power cables and audio cables as much as possible.
2. Strap chassis together. Use heavy battery straps between racks to minimize AC ground potential differences.
3. Use balanced connections if possible. Shielded twisted-pair cable should be used.
4. Transformer-isolate audio inputs if necessary.
5. As a last resort, break the safety ground, but be very careful!

The following figure illustrates the proper use of shielded-twisted-pair cable to reduce ground loops.

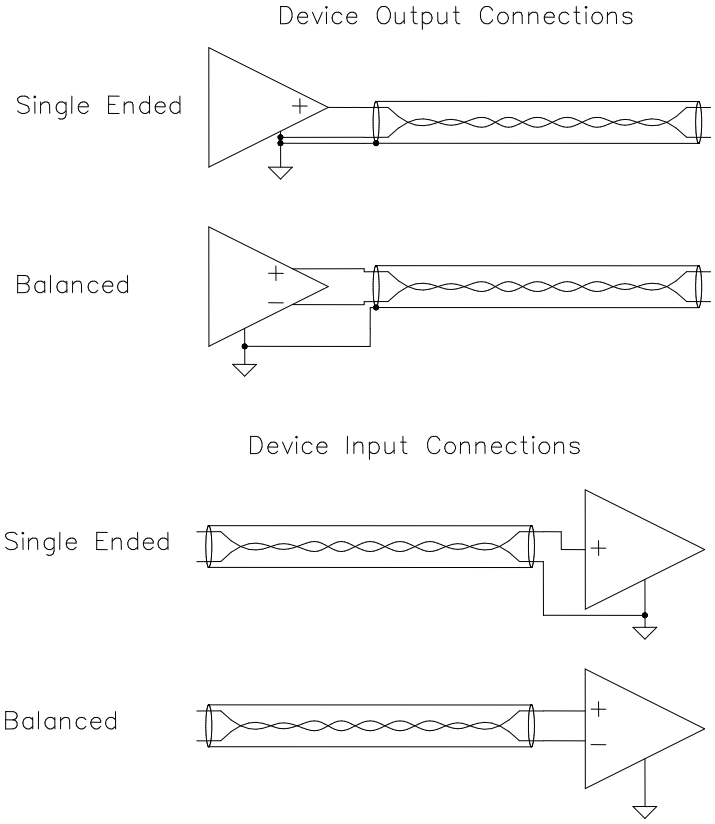


Figure 3-4: Audio Cable Interconnections

NEVER attach the shield at the input device end! All chassis should be strapped together with a separate bus. Do not use the audio cable shield to ground chassis together.

EXERCISES

1. Compute the equivalent resistance of the following series combinations.
   a. 1000 Ω and 250 Ω
   b. 250 Ω and 0.1 Ω
   c. 17 Ω, 23 Ω, 40 Ω, and 20 Ω

2. Compute the equivalent resistance of the following parallel combinations.
   a. 10 Ω and 5 Ω
   b. 100 Ω and 1 Ω
   c. 50 Ω, 60 Ω, 40 Ω, and 100 Ω

4. ACOUSTIC CHARACTERISTICS OF SPEECH

4.1 Speech Production

Speech is the primary process by which humans communicate. The production of speech utilizes the lungs, the vocal cords, vocal tract constrictions, and the oral and nasal cavities. Those parts of the vocal tract that position to form sounds are called *articulators*: the positions of the jaw, tongue, lips, teeth, and soft palate (velum) assist in forming specific sounds. See Figure 4-1. The mouth and nose radiate the acoustic energy.

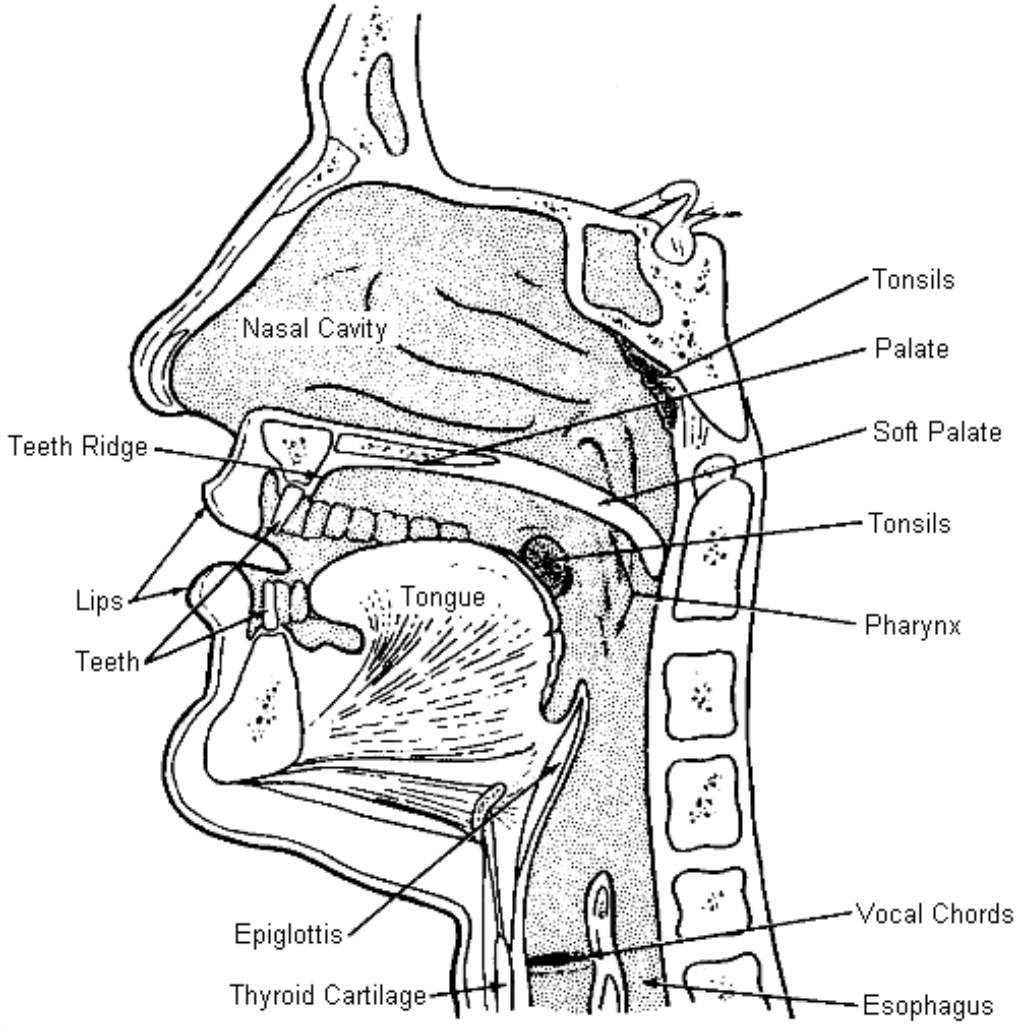


Figure 4-1: Cutaway View of the Vocal Tract

The vocal tract is schematically illustrated below (Figure 4-2).

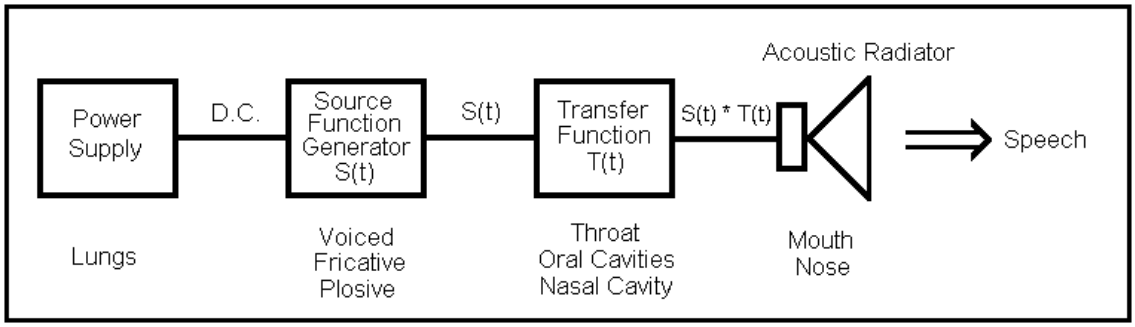


Figure 4-2: Functional Model of Vocal Tract

The vocal tract resembles a clarinet. The air supply from the musician causes a reed to vibrate, which in turn excites a tuned cavity. The buzzing sound is tuned by the resonant cavity and is, in turn, radiated by the flanged horn which couples the instrument's sounds into the air.

These basic speech sounds are called *phonemes*. Table 3 on the following page uses the International Phonetic Alphabet (IPA) to show a list of phonemes commonly used in American English. The vocal tract is capable of producing many more sounds than listed, and different languages use additional phonetic events, *e.g.*, nasalized vowels in French.

One common method of classifying phonemes is by the manner of articulation. At the broadest level, *vowels* allow unrestricted airflow in the vocal tract, while *consonants* restrict airflow at some point. In vowels, such as /i/ in beet, the airstream is relatively unobstructed because the articulators do not come close together. The position of the tongue is of primary importance in producing vowels, though the lips and other articulators have some effect. Consonants are often divided into various subcategories. *Semivowels*, such as the /r/ and /w/ sounds, are vowel-like articulations. In *nasals*, such as the /m/ sound, the air is stopped in the oral cavity but goes out through the nose. *Fricatives*, such as the /s/ and /v/ sounds, excite the vocal tract with air forced through a narrow constriction. *Plosives*, or stops, such as /k/ or /p/, excite the vocal tract with a burst of air.

Another important concept in speech is that of voicing. Speech sounds may be *voiced* or *unvoiced*. In voiced sounds, the vocal cords vibrate and produce pitch. All vowels are voiced, while consonants vary in their voicing. Voiced consonant sounds include the nasals and semivowels. Unvoiced, or voiceless, sounds do not have the melodious pitch of voiced sounds because the vocal cords are apart. Some fricatives and plosives are voiced, such as /v/ and /d/, but others are unvoiced, such as /f/ and /t/. *Cognate pairs,* such as /z/ and /s/ as well as /k/ and /g/, have the same articulation, but one phoneme has vocal cord motion and the other does not. Even normally voiced sounds such as vowels may be intentionally not voiced—by whispering. Try whispering "zip"; it becomes "sip." Laryngitis is an inflammation of the vocal cords and prevents voicing of sounds.

Table 3: American English Phonemes

Phonetic Symbol	Example Word	Phonetic Symbol	Example Word
Simple vowels		**Plosives**	
i	feet	b	bad
I	fit	d	dive
æ	bat	g	give
e	let	p	pot
ʌ	cup	t	toy
ə	uh (unstressed)	k	cat
ɑ	not		
U	book	**Nasal consonants**	
ɔ	law	m	may
u	boot	n	now
ɜ	bird	ŋ	sing
ɚ	Bert	**Fricatives**	
Complex vowels		z	zero
eI	pain	ʒ	vision
oU	go	v	very
aU	house	ð	that
aI	ice	h	hat
Iu	few	f	fat
ɔI	boy	θ	thing
Semivowels and liquids		ʃ	shed
j	you	s	sat
w	we	**Affricates**	
l	late	tʃ	church
r	rate	dʒ	judge

Place of articulation	Manner of articulation				
	Plosive	**Fricative**	**Semi-vowel**	**Liquid**	**Nasal**
Labial	p b		w		m
Labio-Dental		f v			
Dental		θ ð			
Alveolar	t d	s z	j	l r	n
Palatal		ʃ ʒ			
Velar	k g				ŋ
Glottal		h			

The human voice communicates an abundance of information, of which understandable words are but one part, albeit usually the most important part. The three types of information in the voice signal are:

- Intelligibility – what is being said.
- Identification – who is saying it.
- Prosody – the emotional and dialectical aspects of the speaker.

So the identity or the speaker, his emotional state, his gender, his level of education, dialectical influences, and physical attributes are also communicated in the acoustic waveform. In order to preserve this information and assure successful subsequent extraction, the voice should be recorded and processed with sufficient fidelity.

Figure 4-3 illustrates both time and frequency domain representation of a 160 msec interval of speech. The first two time divisions are a low energy fricative which transitions into a higher energy vowel (observe the periodicity). The frequency domain representation (bottom) is dominated by the energy concentration in the vowel).

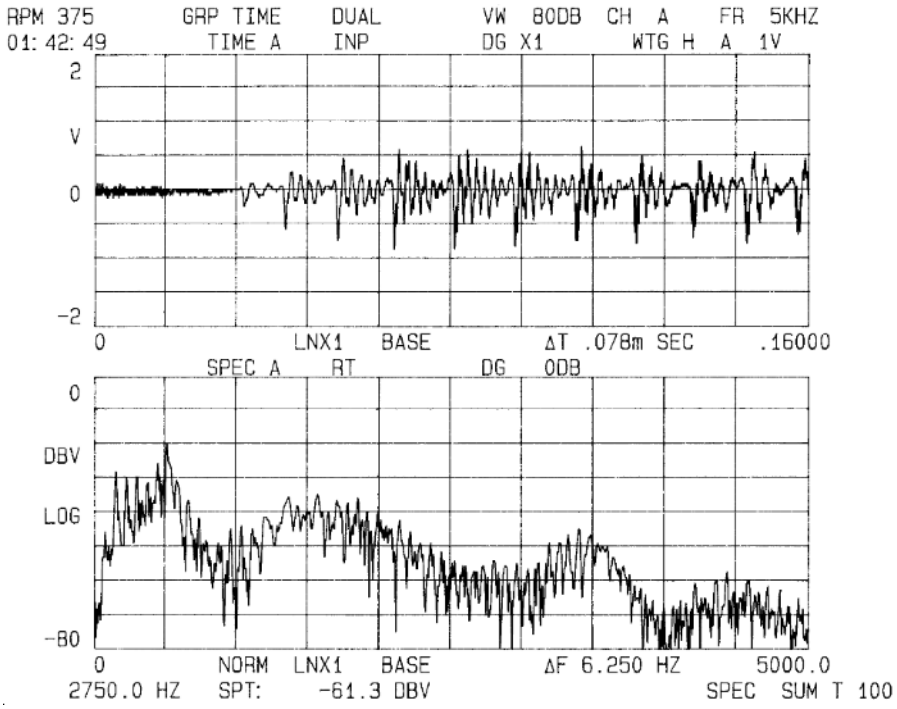


Figure 4-3: Speech Waveform

The fundamental frequency (F0), or pitch, is the repetition rate of a periodic speech signal. Using "eyeball" spectral analysis, one can observe the pitch (repetition) interval of the vowel at approximately 13 msec. The spectrum shows strong spectral lines spaced every 1 / 0.013 s = 77 Hz. This pitch energy corresponds to the Source Function in Figure 4-2.

The overall shape of the spectral envelope shows peak resonances at 500, 1500, 3400, and 4500 Hz. These are called *formants* and are produced by resonances in the vocal tract, *i.e.*, the Transfer Function of Figure 4-2.

Voiced speech sounds are approximately periodic in nature. They each consist of a fundamental frequency (the pitch) component along with harmonics. Depending upon the vocal tract configuration, different harmonics have different amplitudes. The shape of these amplitudes is the *envelope* and is shown in the figure below as a dashed contour.

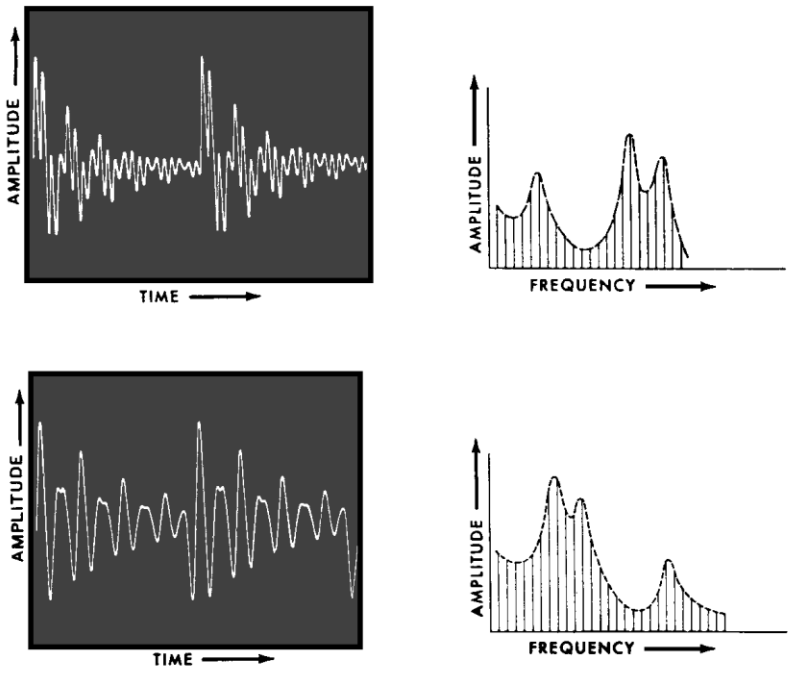


Figure 4-4: Typical Time Waveform &
Corresponding Power Spectra of Voice Speech

4.2 Speech Perception

The ear is the organ of speech perception and is partitioned into three sections: outer ear, middle ear, and inner ear. Figure 4- illustrates.

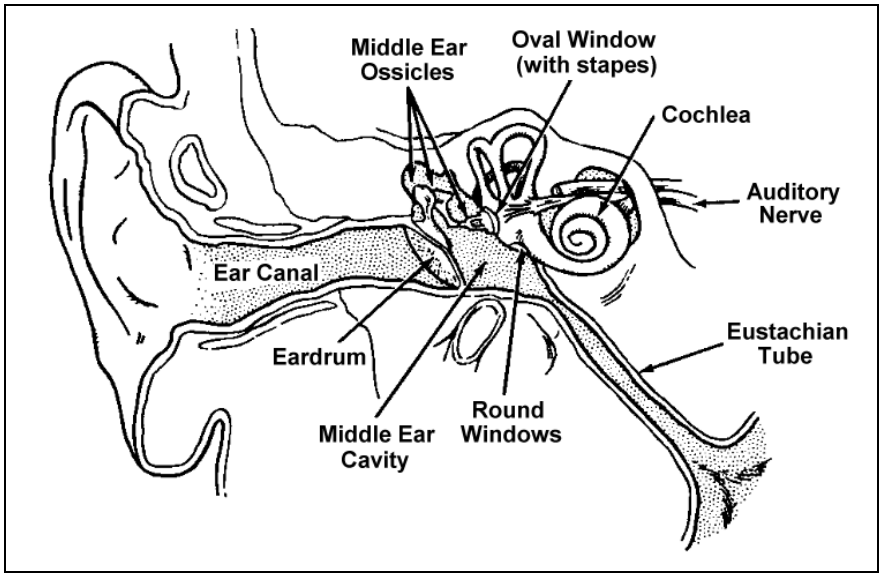


Figure 4-5: Cutaway View of the Ear

The *outer ear* acts as an acoustic amplifier. It consists of the ear funnel, or *pinna*, which directionally collects sounds and the ear canal, or *meatus*, which conducts the sounds to the ear drum, or *tympanum*. The ear funnel is very directional, especially in the middle voice frequencies around 1500 Hz. The ear canal couples the sounds into the eardrum, increasing pressure. This 3 cm canal reaches a peak resonance at frequencies around 3000 Hz. The ear's maximum sensitivity (see Figure 2-1) is partially due to this resonance effect.

The *middle ear* is an impedance-coupling device which couples sounds in less dense air to the very dense fluid of the inner ear. The middle ear consists of a vented (via the Eustachian tube) chamber containing three small ossicle bones: hammer, anvil, and stirrup (mallus, incus, and stapes). These bones connect and give a mechanical advantage of 35 to 80 times. An additional feature of the bone coupling is to provide an automatic gain control operation. The pressure to the inner ear is automatically reduced in the presence of loud sounds.

The *inner ear* is a fluid-filled *cochlea* (coiled, snail-like chamber). The middle ear's stirrup presses on the inner ear's oval window, converting the sound waves to fluid motion. The cochlea, if uncoiled, resembles a long chamber separated for most of its length into two chambers by a membrane. Tiny hair-like cells along this membrane convert motion in the fluid into nerve impulses for processing by the brain's auditory cortex. These hair cells are very fragile and are snapped in the presence of loud sounds, causing permanent hearing loss.

4.2.1 Audio Bandwidth Requirements

Pitch is the perception of the fundamental frequency (F0) of a signal. Men generally have a pitch range around 50–250 Hz, while women have a pitch range of 120–500 Hz. Even when the actual fundamental frequency of a speaker's voice is not present in an audio signal, the presence of harmonics lets us perceive the pitch. Thus, it is not necessary for the lower bandlimit of the speech signal to be below the actual F0 of the speech in order for the pitch information to be conveyed to listeners.

Vowels are characterized by vocal tract resonances known as *formants*. Figure 4- illustrates these frequencies for the vowels. The two principal formants (F1 and F2) have a frequency range of approximately 300 to 3000 Hz. Telephones pass audio in this frequency range, which results in effective voice communication. Speaker identification and fricative discrimination suffers, however, because such information is conveyed at higher frequencies. In addition, intelligibility (understandability) of noisy speech is also impaired by such bandlimiting.

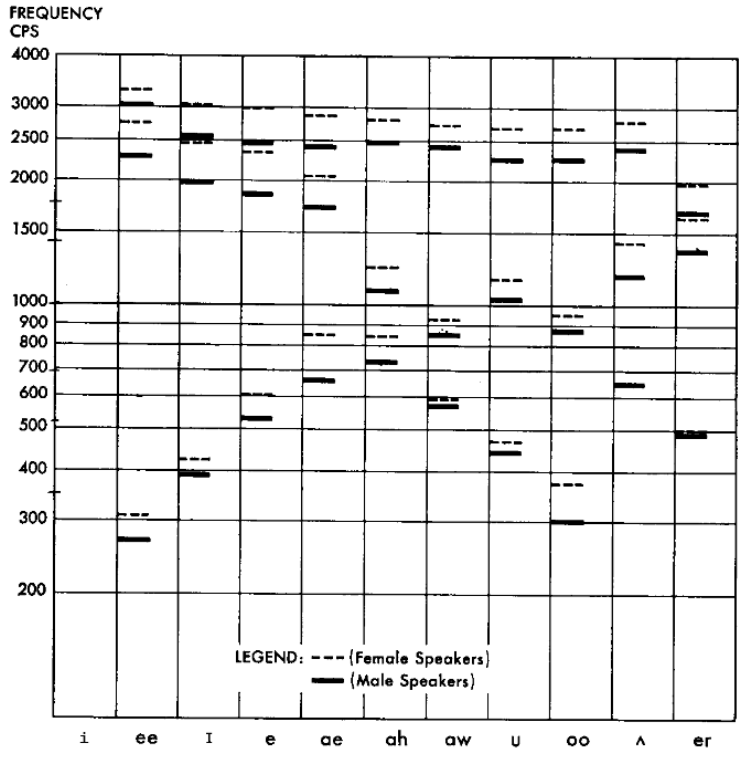


Figure 4-6: Typical Formant Frequencies for Vowel Sounds

Improved voice communication can be achieved by expanding the audio bandwidth. Reducing the lower bandlimit to 200 Hz and increasing the upper limit to 5000 Hz substantially improves voice quality.

Some marginal quality can be achieved by increasing the upper bandlimit to 7000 Hz. The signal-to-noise ratio (SNR) may actually be lowered for a speech signal when noise is present. SNR is defined as

$$\text{SNR} = \frac{\text{signal power}}{\text{noise power}} .$$

Many processes distort the speech waveform enough that the SNR is not a meaningful perceptual measure. Thus, speech processing engineers often use other objective quality measures such as the Itakura-Saito measure (ISM) and weighted-spectral slope measure (WSSM). Still, the SNR and SNR-based measures are appropriate for the form of audio enhancement used for forensics.

Since voice energy greatly drops in energy at higher frequencies (See Figure 2-3), high frequency noise energy may mask high frequency noise speech components. As a result, increasing bandwidth may actually result in admitting more noise than speech.

Though measurable voice energy exists well above 10 kHz, little benefit can be realized in utilizing these higher frequencies due to the potential loss from high frequency noise and minimal gain in voice quality.

4.2.2 Perceptual Phenomena

Nonlinearity – The ear is highly nonlinear and produces harmonic energy due to distortion. The level of distortion is difficult to measure. As an example, a 200 Hz tone in one ear and 420 Hz tone in the other ear will produce a 20 Hz perceptual beat, which implies that a second harmonic has been produced from nonlinear distortion.

Phase Sensitivity – Experts disagree as to the ear's phase sensitivity; however, it appears that the ear is relatively insensitive to minor (<100 msec) phase shifts.

Loudness Sensitivity – For simple sounds such as tones, the ear's loudness response appears to be very logarithmic. A 10 dB SPL increase appears to correspond to a doubling of perceived sound volume. Perceptual volume is measured in *sones*. The ear appears to be more discriminate at low audio levels, suggesting that careful listening should be conducted in a quiet environment at low audio levels.

Frequency Sensitivity – The ear's ability to discriminate different frequency tones is similarly nonlinear. The ear appears to be octave based, *i.e.*, a doubling in frequency represents a fixed perceived pitch change. Frequency resolution is significantly better at low frequencies than high. Perceived pitch is measured in *mels*.

Directionality – The direction of a sound source to the head is primarily determined by the difference in sound arrival times to the two ears, *i.e.*, delay differences. Since the ears are

separated by approximately 20 cm, differential sound arrival delays of 500 μsec are significant (sound travels at 34,400 cm/sec as per Table 1).

A secondary effect aiding sound direction is the spectral difference in the sound due to diffraction effect of sound bending around the head and the effect of sound reflections from the skin founds in the ear funnel.

Reverberations and Echoes – The ear seems to interpret sound reflection arriving less than 30 msec after the direct sound as a reverberation. This is referred to the *presence effect*. The 30 msec zone is known as the *fusion zone*. Reflections arriving later than 100 msec are perceived by the ears as a discrete echo. The 30 to 100 msec zone is a transition region in which neither echoes nor reverberations are distinctly discernible.

The overall process of delay zones is known as the *Haas effect*. The fusion zone may be extended by sufficiently loud reflections arriving within successive 30 msec time intervals. For this reason, multiple, closely-spaced reflections produce long reverberation times.

4.3 Voice Identification

Forensic voice identification, often called "voice printing," utilizes both aural (listened) and spectrographic information. A voice spectrogram utilized in such examinations is illustrated in Figure 4-. The horizontal time base is approximately 2.5 seconds, and the vertical frequency range is 4 kHz. The dark arched bands are vocal tract formants and the vertical striations result from the vocal cords (glottis) opening and closing. Note the lack of such voicing with the "S" sound. The voicing and energy are intense in the "A----W" segment.

Spectrograms can be used to recognize and identify speakers. Some experts can match reference spectrograms to test spectrograms by the same speaker, but a disguised voice can be difficult to match. Experience has shown that certain voices are more difficult to identify via spectrograms than others. Though there is evidence of reliable use of voiceprints, many researchers feel that spectrogram reading alone has not been shown to be a reliable form of speaker recognition. Despite the controversy about voiceprints, several courts of law in the United States, Canada, and Europe have admitted them as evidence.

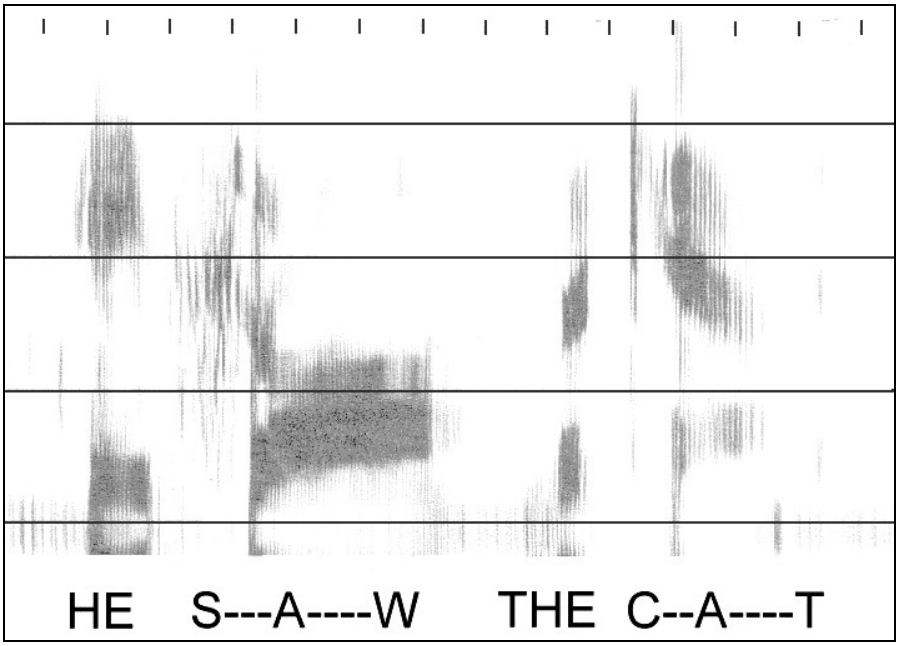


Figure 4-7: Voice Spectrogram

Automatic speaker recognition can be performed by computers executing speech processing algorithms. In the problem of *speaker identification*, the computer is given speech from an unknown speaker and must determine which of several known speakers produced the speech. In the *speaker verification* task, the computer determines whether the given speech was produced by the expected individual or by someone else.

EXERCISES

1. If speech is bandlimited to 2000 Hz, which sounds would suffer most, vowels or fricatives?

2. What is the vocal tract's source function during whispered speech? Where is this energy produced?

3. What vowel typically has the highest second formant?
   The lowest second formant?
   What are their average frequencies in males?

4. The phonetic transcription for "feet" is /fit/.
   What are the phonetic transcriptions for the following words?
   a. "bat"
   b. "shoot"
   c. "slow"

5. NOISES AND EFFECTS

At the most basic level, the two factors that can make voices unintelligible are collectively known as *noises* and *effects.* These are discussed in more detail in the following subsections.

5.1 Noise/Effect Model

Intelligibility degraders fall into two categories according to the means by which they are produced:

Noises are produced from energy-radiating sources which add to the desired sound or audio, which is usually one or more voices. Background music, machinery noise, traffic noise, and 60 Hz power hum are examples of noises. Simply subtracting noises from the signal will better reveal the voices, thereby improving intelligibility.

Effects are properties of the environment and the equipment that modify the sound or audio whenever it is present; these must be systematically reversed to improve intelligibility. Examples of effects include echoes, reverberations, hidden microphone "muffling", near party / far party, and *non-linear distortion.* Effects, particularly acoustic echoes and reverberations, are extremely complex; the fact that effects are applied to not only the desired voices, but also to the added noises, further complicates the issue. Generally, complex mathematical models must be developed in order to reverse such effects and improve intelligibility.

A noise/effect model is shown in Figure 5-1. A noise source (cloud) and a voice combine acoustically. This combined sound is modified by the acoustic effects of the room before the microphone picks it up and converts it to audio. The audio is then further corrupted by the noises and effects of the transmission channel that is used, *e.g.*, a telephone line or RF link.

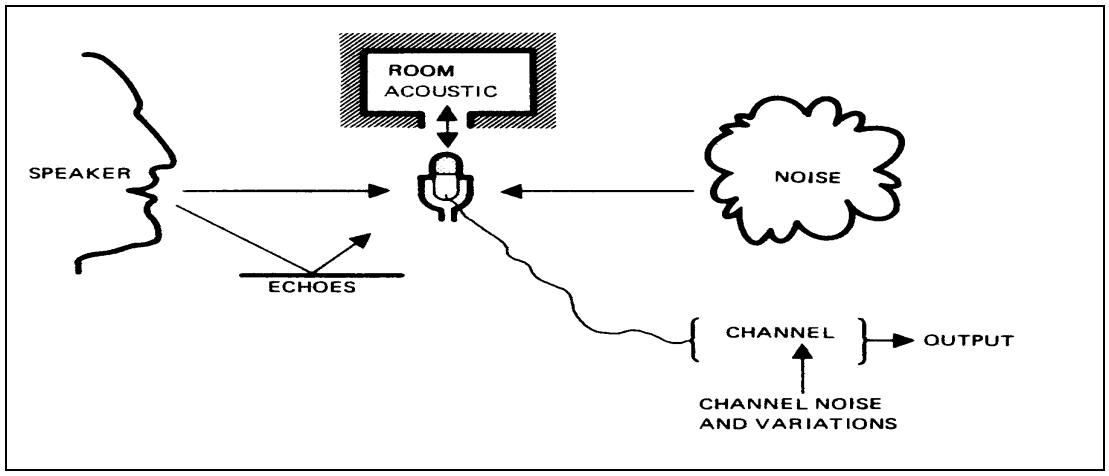


Figure 5-1: Noise/Effect Model

5.2 Types of Noises

Noises may be broadly classified as either time-correlated ("predictable") or non-time-correlated ("random"). This property is particularly important in tape enhancement adaptive filtering, discussed later. Such filtering distinguishes between voices and noises using a signal prediction procedure.

Audio signals are always a weighted blend of predictable and random signal components; some signals can be more predictable or random than others, and vice-versa. There are *degrees* of predictability, as well as degrees of randomness, so it is not a simple black-and-white matter to separate the two signal types from a complex signal.

A high degree of *autocorrelation* in a signal (comparing it with itself over time) implies that the signal's waveform in the future can be predicted by observing the waveform's history. In fact, we say that such a signal is predictable and that a signal with a low autocorrelation value is random.

The best example of a highly correlated waveform is a sine wave (tone), as shown in Figure 5-2. The sine wave is a simple waveform that repeats over and over again, *ad infinitum*. Each *period*

Figure 5-2: Sine Wave (Time-Domain View)

(360° or 2π radians), or cycle, is a copy of a previous period. The signal possesses a high degree of autocorrelation, and so by observing the past, the future of the waveform can be very accurately predicted.

On the other hand, random signals have little or no autocorrelation from one time segment to the next. The history of a random signal thus yields little or no information about its future waveform. Purely random signals are completely uncorrelated, and hence are not predictable at all; the most extreme example of a random signal is *white noise*, as shown in Figure 5-3.

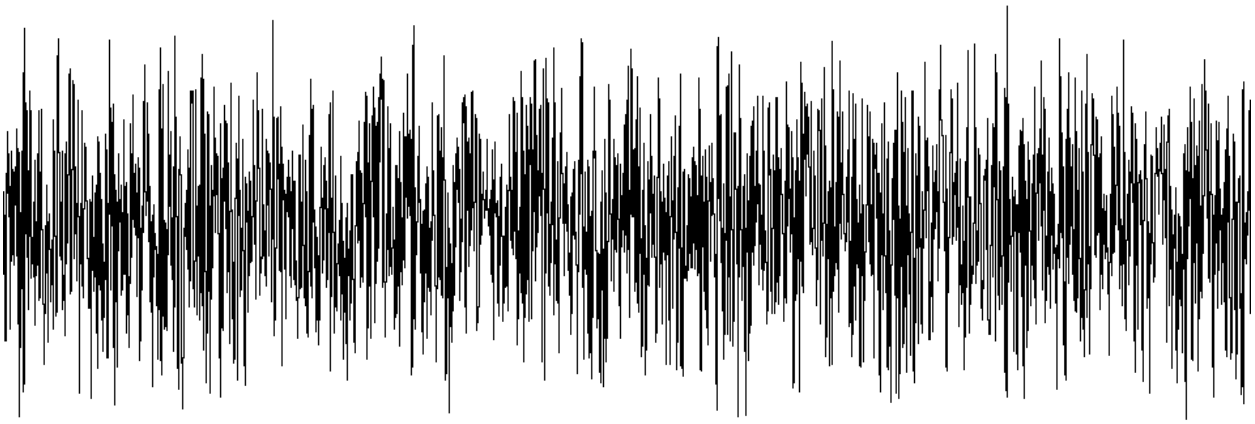


Figure 5-3: White Noise (Time-Domain View)

As pointed out earlier, the power spectrum of a sine wave is a single line, *i.e.*, all energy is concentrated at one frequency, as shown in Figure 5-4. Such a noise is easily reduced, simply by suppressing just that one frequency in the signal, with little or no effect on broader bandwidth voice signals that may be present.

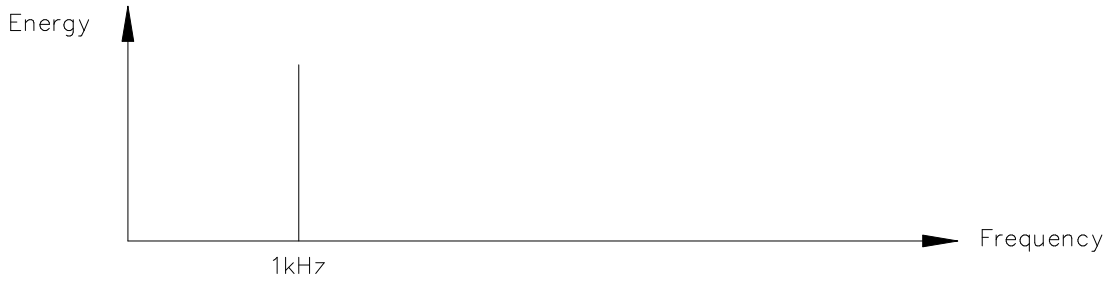


Figure 5-4: Power Spectrum of 1 kHz Sine Wave

Random signals, on the other hand, have their power dispersed over a broad range of frequencies, including those containing the voice signal. The power spectrum of white noise is flat with equal energy at all frequencies. Figure 5-5 illustrates the power spectrum of white noise. Successfully reducing such a noise without severely degrading an underlying voice signal is extremely problematic, though there are various methods available, mostly variations of a technique called

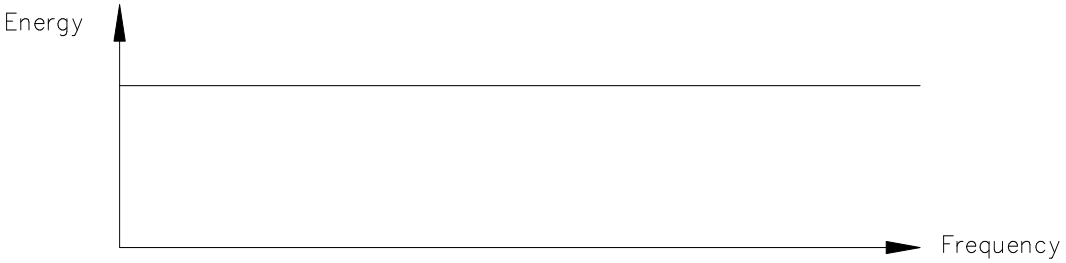


Figure 5-5: Power Spectrum of White Noise

spectral subtraction, which will be discussed more later.

Few additive noises are purely correlated or purely random. Generally, they fall into the following three categories:

> Banded Noise – Banded noises are those whose energy occurs at distinct frequencies and/or frequency ranges. Some banded noises, such as AC "mains" hum, are highly predictable, though there are usually random components as well. Other banded noises, such as analog tape "hiss", are concentrated at high frequencies. Room noise and rumble, as well as automobile engine noises, are often concentrated in low frequencies below 200 Hz. Banded noises may be effectively attenuated using conventional "surgical" techniques—such as comb, multiple notch, bandpass, lowpass, and highpass filtering—as long as the noise spectrum does not overlap the voice spectrum excessively.

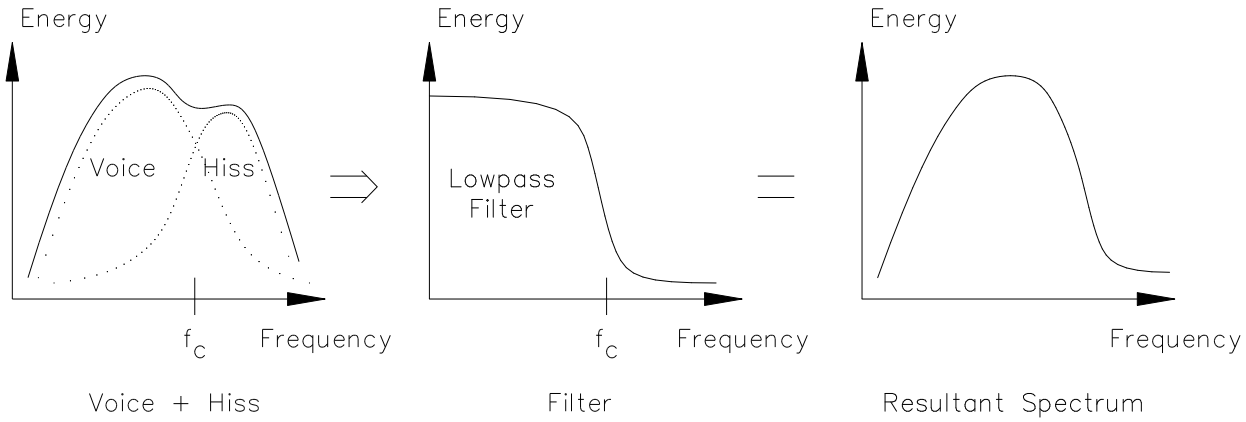


Figure 5-6: Banded Noise Spectrum - Tape "Hiss"

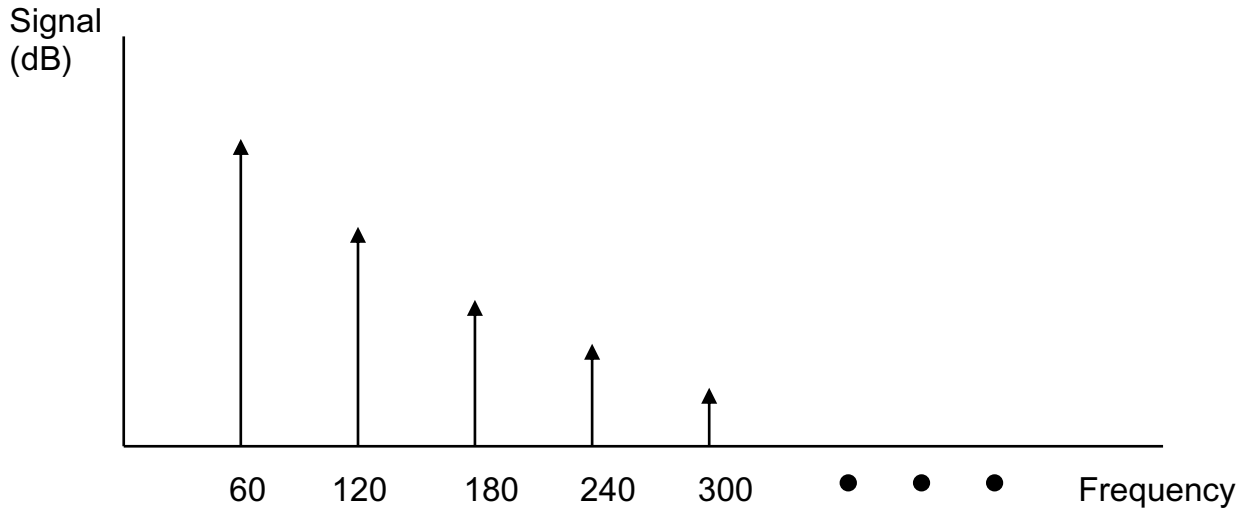


Figure 5-7: Banded Noise Spectrum - AC "Mains" Hum

Predictable Broadband Noise – Broadband noises have excessive overlap with the voice spectrum, and thus cannot be effectively treated with techniques that suppress specific frequencies in the way that banded noises can be. Therefore, more sophisticated approaches are required. A special case of broadband noise, *predictable* broadband noise, can be treated using adaptive filtering techniques that exploit the relative high degree of predictability of the broadband noise signal as compared with that of the voice signal. Consider, for example, any type of machinery or motor noise; such a noise signal generally consists of a broadband signature (as shown in Figure 5-8) that repeats over and over with time, due to a motor operating at a relatively constant speed and making essentially the same sound on each rotation. Given a sufficiently long analysis window, an adaptive filter operating in the *time domain* can identify this *time-correlated* noise and effectively attenuate it, without significantly degrading the voice signal. Certain types of background music can also be effectively reduced by such adaptive filtering, given the harmonically-structured notes which have a high degree of predictability. The adaptive nature of the filter means that speed and note variations can be automatically tracked to maintain optimal cancellation at all times, which means that no operator intervention is required once the filter is setup and running.

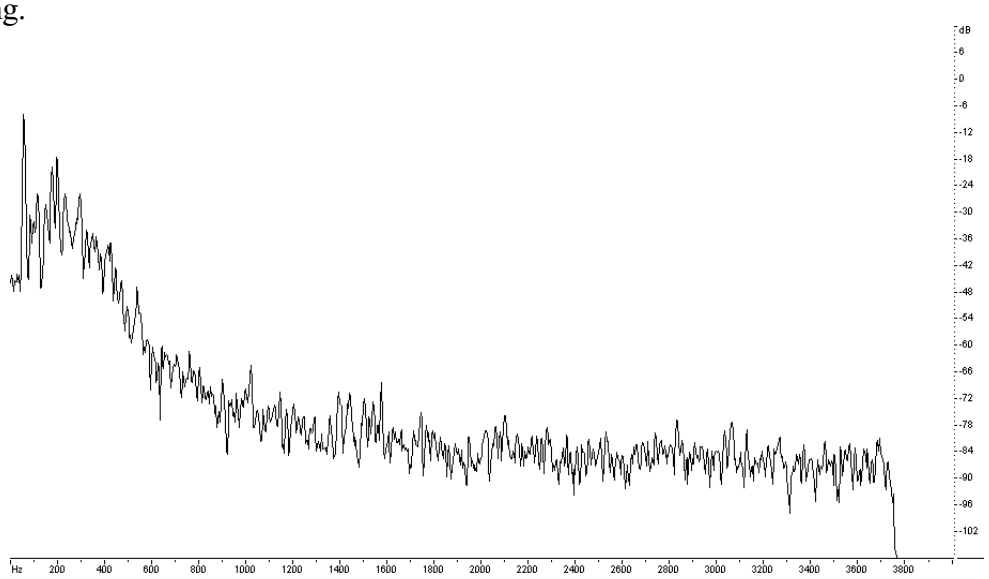


Figure 5-8: Broadband Noise Spectrum

Random Broadband Noise – Broadband noises that are not time-correlated (never repeat) are not predictable, and are referred to as *random* broadband noises. Examples of such noises include RF static, airflow noises (e.g. wind), and waterflow noises (e.g. ocean surf, shower, etc.). Time-domain adaptive filtering techniques generally have no effect on reducing these types of noises, because there is no repeating time pattern to identify. However, in the frequency domain it is possible to identify the spectral signature of the noise and, using a technique known as *spectral subtraction,* reduce the noise signal without reducing the voice signal. Automatic spectral subtraction filters, such as the ones found in the DAC PCAP and other products, work by updating a *noise estimate* profile whenever voices are not present in the signal, then continuously subtracting this profile from the input

signal even when voices are present. Noise reduction results can be quite dramatic with such filters, *as long as the voices are significantly louder than the noise to begin with and the noise is relatively constant*. If the voices are buried in the broadband noise, or the spectral characteristics of the noise are constantly changing, degradation of the voice signal will generally occur, due to the voices being mistaken by the algorithm as noise and applied to the noise estimate. Also, automatic frequency-domain filters, of any type, will generally introduce "birdy-noise" artifacts into the signal output due to the block-based (not sample-by-sample) updating of the filter solution.

5.3 Types of Effects

5.3.1 Environmental (Acoustic) Effects

Acoustic effects often degrade the intelligibility of voice sounds within by modifying their normal characteristics. For example, the *reverberations* of a voice within a room will cause the loud sounds in the voice, the vowels, to mask subsequent weaker sounds, the consonants, in the voice, degrading intelligibility. This is illustrated in Figure 5-9.

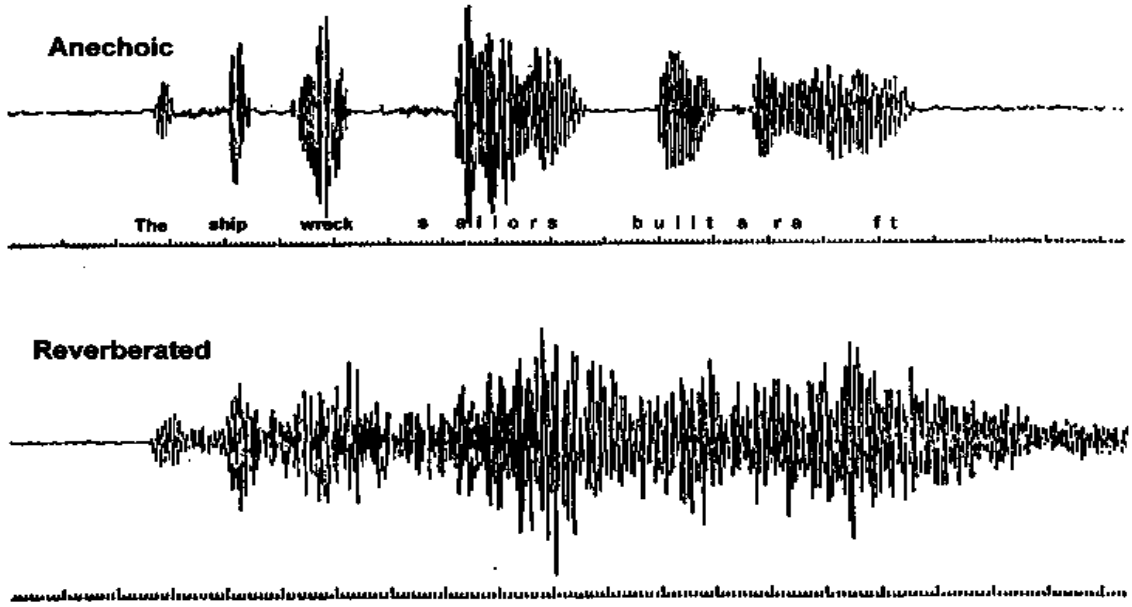


Figure 5-9: Anechoic vs. Reverberated Speech (Time-Domain View)

Fortunately, such acoustic effects can be predicted and cancelled by an adaptive filter, due to the fact that the desired signal, in this case the desired voice, is a repeating pattern due to the reverberation. As the filter operates, the room effects will be recognized and systematically reversed, bringing the filtered output closer and closer to the original, anechoic, sound.

Another common environmental effect is referred to as *muffling.* As illustrated in Figure 5-10, the high-frequency portion of the voice sounds is often attenuated due to *absorption losses* in the materials through which the sounds must pass to get to the microphone, e.g. clothing and/or foam. Or, the recording equipment may inherently have a severe high-frequency rolloff characteristic, due to being low-quality or designed extended record time at the expense of bandwidth.

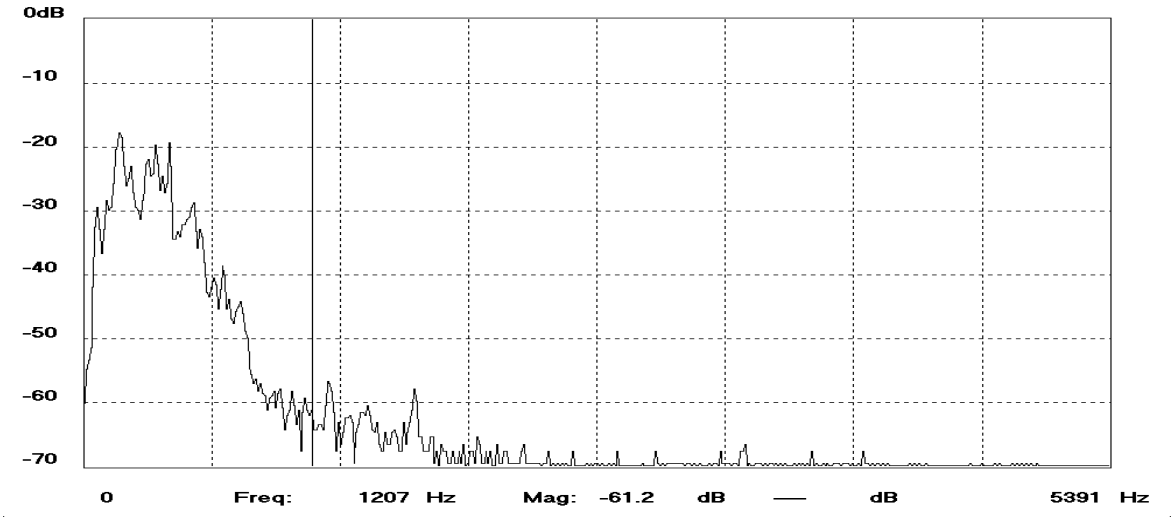


Figure 5-10: Muffled Voice Spectrum

As illustrated in Figure 5-11, muffling can be largely corrected in most cases, simply by applying *equalization* to the input signal to bring up the high-frequency sounds. Although high-frequency background noises will also be increased by the equalization, voice intelligibility will often be improved; this is a prime example of how we sometimes have to make the audio "sound worse" in order to make the voices more understandable.

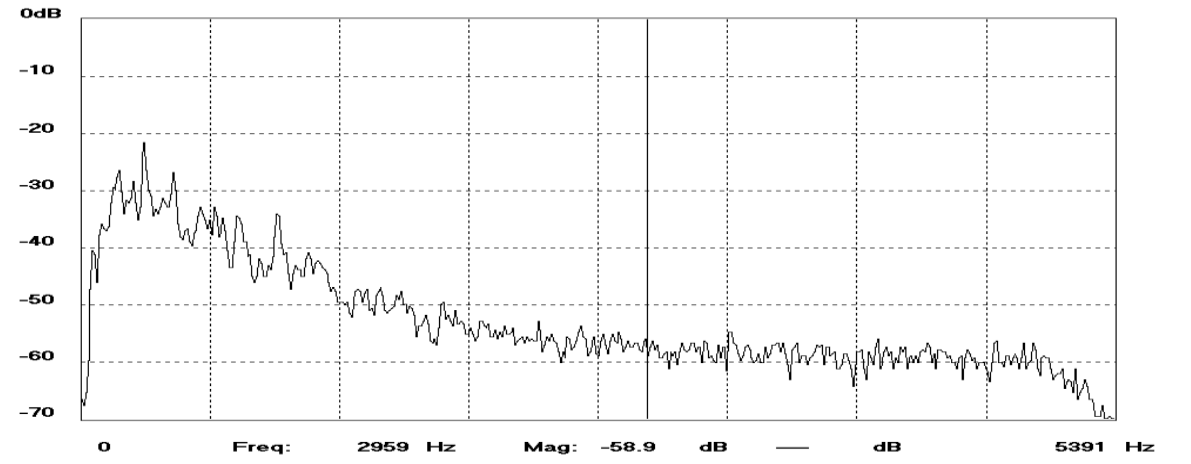


Figure 5-11: Muffling Corrected by Equalization

Another common acoustic effect is *near/far party.* This occurs whenever one person in the conversation is close to the microphone and the other is farther away. Naturally, the person close-up will be louder, and the other person will be quieter. When trying to understand both sides of the conversation, this level difference makes understanding the far party difficult, even if the recording is otherwise clear and free of background noises. Therefore, *dynamic level processing* of the signal is often employed to bring both sides of the conversation to similar levels.

Figure 5-12: Near/Far Party

5.3.2 Equipment Effects

As with the environment where the sound is occurring, once the microphone has converted the sound to audio, additional effects can be imparted to the voice signal, rendering it less intelligible. For example, consider a trans-oceanic telephone call; often, when you do not have a good-quality line, whenever one person speaks he hears himself again a short time later; at the least, this can be extremely annoying, and at worst it makes normal conversation impossible. In this case, what is happening is that the telephone equipment is creating the echo effect, in a very similar manner to the way that acoustic reverberations occur within a room; in fact, the very same adaptive filters that are used to reduce reverberation are also used to reduce telephone line echo. In the latter case, we refer to the filter as an *echo canceller.*

Muffling effects can also be introduced by the equipment after the microphone has picked up the sound. For example, a microcassette recorder has extremely limited audio bandwidth, due to the slow speed of the tape and the corresponding limited ability of the tape to store magnetism at higher frequencies without becoming *saturated*. Again, equalization can be used to correct this.

But the worst issue related to equipment is called *non-linear distortion.* Such distortion occurs whenever signal information is irreversably lost by the equipment. Examples of non-linear distortion include “audio clipping” due to overdriven or poor-quality microphone and/or limited power supply voltage on equipment, tape “dropouts” due to poor-quality media being used and/or

dirty record heads, and "lossy" digital compression such as that used by minidisc and MP3 recording devices to reduce bit storage requirements and thereby extend record times.

Of these equipment effects, however, the audio clipping phenomenon is one that has at least some chance of being treated effectively; consider the clipping of a 1kHz sine wave as illustrated in Figure 5-13:

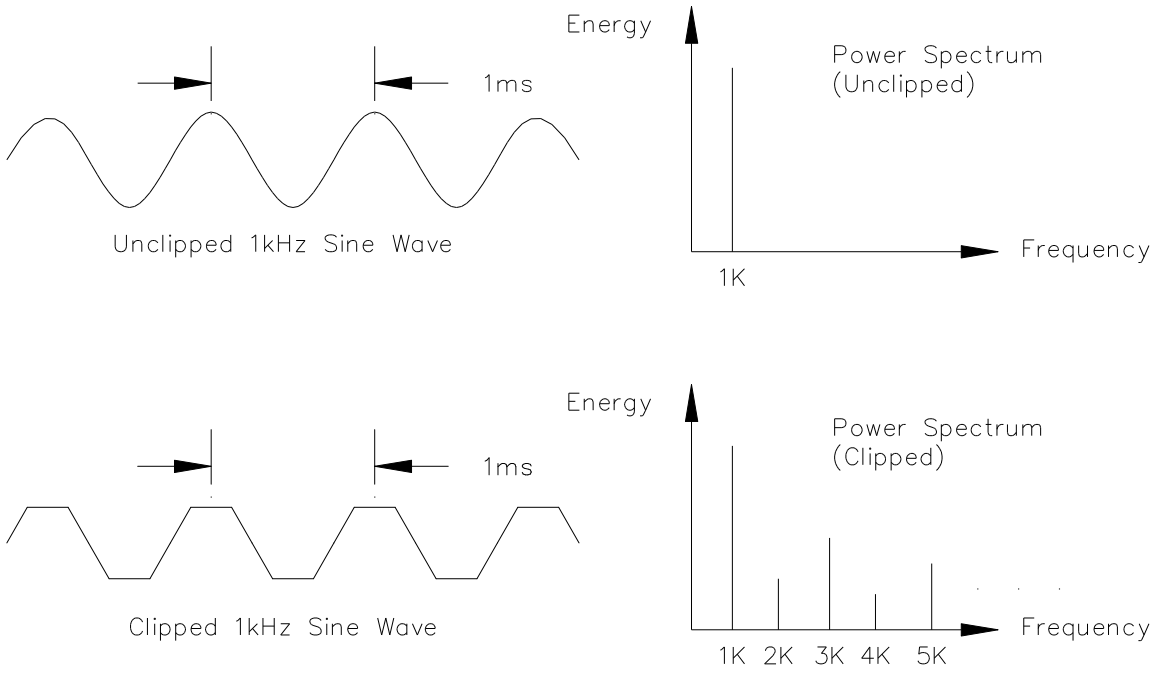


Figure 5-13: Sine Wave and Clipped Form

Looking at the clipped waveform, it is obvious that if the "flattened" peaks could be made more "rounded", much of the adverse effect of the clipping would be reduced. From a forensic standpoint this creates a bit of a dilemma, in that whatever process that is being used to re-round the peaks could be said to be adding new signal information to the recording that was not present originally.

But from a frequency-domain perspective, again illustrated in Figure 5-13, re-rounding the peaks would effectively attenuate the additional frequency harmonics that were created by the clipping. These are known as *harmonic distortion components,* which occur at multiples of the sine wave frequency (2000, 3000, 4000 Hz, etc.). Thus, we can legitimately argue that re-rounding the clipped peaks, if properly done, would not add anything to the original audio signal, but would simply reduce the unwanted harmonic distortion that was produced by the clipping.

Clipped peak restoration is a technique that can be used to re-round flattened signal peaks, and is discussed later in this manual in Chapter 12.

EXERCISES

1. Identify which of the following are noises and which are effects:
   a. music
   b. AC hum
   c. room reverberation
   d. telephone crosstalk
   e. echoes

2. An analyzer makes the following noise power measurements on a voice recording:

Freq. Band	Measured Power
100–200 Hz	1.0 watt
200–400	1.0
400–800	0.5
800–1600	0.2
1600–3200	0.01

   With nothing else to go on, what would you guess is the problem with this recording?

3. A loud 1.5 kHz tone overdrives a microphone circuit. What frequency(s) do you expect to see in the output audio?

4. A microphone is placed near an air conditioner unit, where it picks up both the sound of the air conditioner running and the sound of the airflow. Which of these sounds is random, and which is predictable?

5. Sound absorption material is introduced into a room. Would you expect this to increase of decrease the amount of reverberation that occurs? Why? If the absorption material covered up a microphone previously installed, how would you expect the output audio to be affected?

6. ACOUSTIC CHARACTERISTICS OF FORENSIC ENVIRONMENTS

6.1 The Real World, Not Hollywood…

Forensic audio differs significantly from professional audio in several areas, including:

- Covert, non-optimal microphone (mic) placement,
- Non-ideal, compact recording equipment and audio transmitters,
- Talkers not cooperating, who may not know they are being recorded,
- Lack of environmental sound treatment as in a studio, and
- No control of background, or "ambient", noises.

In all categories, the forensic technician is highly disadvantaged, and does not have the benefit of the same obvious, almost "no-brainer" solutions as those who produce audio material for the commercial broadcast and music industries. For example, in the forensic world one cannot simply do "another take" and move the microphone closer to the talkers, nor can one bring the talkers back to the studio and use ADR (automatic dialog replacement) equipment to re-record their voices if they are not usable. Also, we can't tell the people talking in the background to shut-up, nor can we simply turn off noisy equipment we hear operating in the background just because these things are making our lives difficult. Add to this the critical value of the audio from the perspective of a criminal prosecution or a national security purpose, and you begin to get a small sense of what it's like to be in the forensic technician's shoes relative to those of the average disc jockey or record producer.

So the job is *much* harder, and is further exacerbated by the lack of direct control of how the recording was made in the first place; in many agencies, for example, one group is responsible for making recordings while yet another group is responsible for processing those recordings. Though this "firewalling" of functions is generally done for good reasons, it means that recordings are often made by operators who don't fully appreciate the strengths and weaknesses of their equipment in terms of what can and cannot be achieved by "the lab geeks". Opportunities to maximize the quality of the end product are therefore often missed.

In recent years, the television program "CSI", as well as the ongoing debacles regarding "the Nixon tapes", has made such problems even worse by creating the false impression, both with lay people and even with some law-enforcement and legal professionals, that any bad recording can be salvaged with sophisticated modern computer technology. Nothing can be farther from the truth; the fact is that modern audio recording and processing technology has both its advantages and its pitfalls, and some of these pitfalls are actually a step backward from the "old days" of analog.

For example, modern digital recording equipment does eliminate many of the equipment-related issues that occurred in the old analog gear, such as wow-and-flutter distortion, misaligned tape heads, tape hiss, etc. If the recorder utilizes an *uncompressed* data format for storage of the

recording, as is the case with CD and DAT media and PCM WAV file-based media, theoretically all the speech-band audio information is right there, waiting to be recovered by state-of-the-art processing equipment, algorithms, and techniques.

But unfortunately, many modern digital recorders that are commonly used employ "lossy" compression technology, such as ATRAC, MP3, and ADPCM formats, in order to extend the maximum record time of the device and thereby reduce manufacturing costs, *at the sacrifice of low-level information* in the audio signal. Minidisc and pocket solid-state dictation recorders are the worst examples of such recorders; these devices, when used in the same applications as older analog recorders, such as microcassette machines, often yield less-usable recordings than the older technology, despite the perception that they are inherently "better, because they are digital".

Such little-understood equipment issues make the job of the forensic technician that much harder today, because now it's more important than ever to recover every bit of usable signal that is present in the recording, knowing that much of "what we could have gotten, if only.." may have been permanently discarded by the recording device as the tape, disk, or file was being made.

Even assuming that the ideal, uncompressed format, digital recording device is used, the workload for the forensic technician has, paradoxically, been *increased* by high-fidelity digital recording technology. Although we might previously have believed that "going digital" would eliminate much of the need for audio post-processing, the truth is that we are now using digital recording devices in applications that previously would never have been attempted with analog equipment, and so the need to deal with extreme environmental noises and effects and recover usable speech product is even more in demand. Plus, with the increased capacities of the digital recording devices relative to tape machines, the sheer magnitude of the collected data that now needs to be processed can often be overwhelming.

Therefore, it is now more important than ever for the forensic technician to fully understand the acoustic characteristics of the forensic environment where the recording is made; by fully appreciating what is happening to the sounds before they are picked up by the microphone, converted to audio, and stored onto the tape, disk, or data file can we achieve the best results in the forensic audio laboratory, and perhaps come closer to being the "miracle workers" that we are often expected to be.

6.2 D-R-A-T

In acoustic environments, all energy from sound waves is subject to the following effects, which we can remember by the acronym D-R-A-T:

Distance, or Inverse Square Law – If a sound wave of power *P* radiates omni-directionally outward from a sound "point" source in an unobstructed manner, *i.e.*, no reflecting or absorbing surfaces, then it will produce a *spherical* wave front. This means that the intensity *I*, measured in watts/meter^2, will decrease as the square of the distance *r* from the source, *i.e.*,

$$I \propto 1 / r^2$$

or more precisely,

$$I = \frac{P}{4 \pi r^2}$$

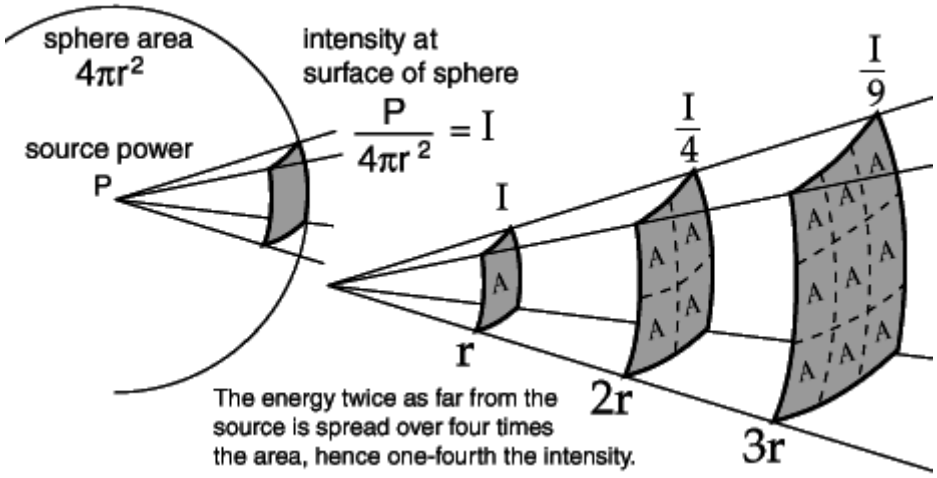


Figure 6-1: Effect of Distance on Sound Intensity

Each doubling of the distance from the source reduces the sound intensity by a factor of four, or 6dB. In a theoretical *free-space* environment, in which a sound is free to propagate forever without ever encountering an obstacle, sound intensity is purely a function of $1/r^2$. But in the real world, the closest one ever comes to free space is outdoors, in an open field with no buildings around. So in most cases, the actual sound intensity will not directly follow the inverse square law, but will be influenced by some degree by the other factors listed in this section.

Reflection – When a sound wave strikes any material, some of the acoustic energy will be reflected. Harder materials tend to be the best reflectors; smooth materials tend to reflect in a single direction, rougher materials tend to scatter the sound in multiple directions.

Absorption – Soft materials will absorb a portion of the sound energy, as indicated by the *absorption coefficient* for the material *as a function of the sound frequency*. For example, acoustic tile with smaller texture dimensions absorbs higher frequencies (shorter wavelengths) better than lower frequencies, and thus has a larger absorption coefficient at higher frequencies.

Transmission – Any sound energy that is not reflected or absorbed by the material will pass through to the other side, possibly changing direction of travel due to different speed of sound in the material and scattering effects.

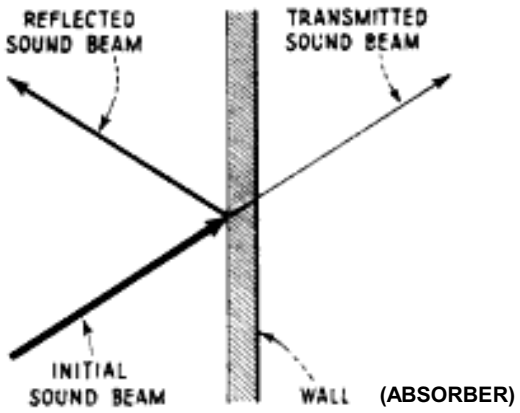


Figure 6-2: Effects of Reflection, Absorption, and Transmission

6.3 Sound Fields

The inverse square law holds true for the **Far field**, which is open space (approximating free space), but enclosed spaces always behave differently.

Sound that is perceived (or picked up, in the case of a microphone) at any point in a room is a combination of direct sounds (if a direct line-of-sight is present) from all sources and reflected sounds from walls, furniture, etc.

If all the surfaces in the room are very good at reflecting sound, the sound field in that room will be **diffuse**, meaning that sounds can be heard fairly evenly at all points. Most rooms, however, will have sonic "shadows", meaning that there are points at which sounds cannot be picked up as well as they can be in others.

Most rooms have 3 sound fields:

- **Near field** - within 1 wavelength of the lowest frequency of a sound produced by a source, SPL measurements vary widely; for human male voice, this distance is approximately 11 feet, or 3.4 meters.
- **Far (Direct) field** – line-of-sight portion of far field between near and reverberant field where inverse square law applies; SPL for source of interest decreases 6 dB for each doubling of distance.
- **Far (Reverberant) field** - near large obstructions, e.g. walls, reverberant (diffuse) sound predominates and the SPL level of the source of interest is relatively constant.

Generally, a microphone with good far-field sensitivity can be placed in either the direct field or the near field and be expected to pickup of the desired source, along with any other noise sources whose sounds are present at that point. However, near-field microphones, such as those used by cellular and landline telephones and many public-address systems, are specifically designed to

attenuate far-field sounds, for obvious reasons; therefore, they must always be within the near field of the desired source in order to have good pickup.

As illustrated in Figure 6-3, far-field omnidirectional, and even cardioid, microphones placed in the reverberant field of a sound source will invariably suffer some reduced intelligibility of the desired source due to the presence of strong reverberations. In the case of cardioid-pattern microphones, which are the most commonly used for audio surveillance, the "null" in the pattern is completely ineffective for reducing undesired sounds and reflections. This is where audio filtering devices that are able to reduce the effects of the reverberation will be of most value.

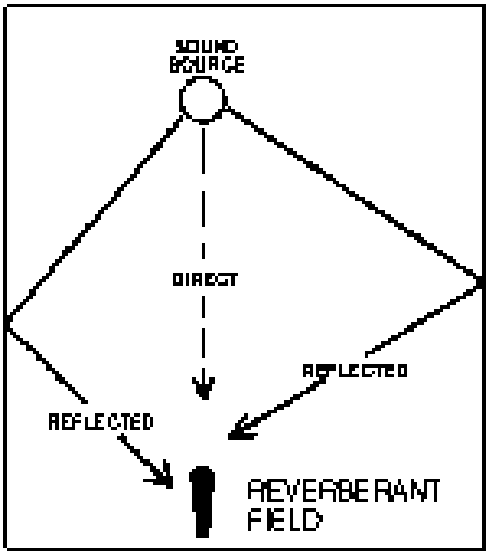


Figure 6-3: Far (Reverberant) Field Mic Placement

6.4 Room Effects

6.4.1 Absorption Coefficient

The absorption coefficient of a material indicates the portion of any sound that strikes it (at a given frequency) that is absorbed by the material and not allowed to pass through. Essentially, it expresses how much of the sound energy is converted to heat within the material.

The absorption coefficient can be expressed as:

$$\alpha = I_a / I_i$$

where

I_a = sound intensity absorbed (W/m^2)

I_i = incident sound intensity (W/m^2)

A perfect sound absorber would have an absorption coefficient of 1.0, which is considered to be analogous to an "open window" in the room. Conversely, a perfect sound reflector would have an absorption coefficient of 0.

Approximate absorption coefficient values for common materials at 500-1000 Hz (which are of most interest for speech considerations) are expressed in Table 4.

Table 4: Absorption Coefficient Values for Typical Building Materials (500-1000 Hz)

Plaster walls	0.01 - 0.03
Unpainted brickwork	0.02 - 0.05
Painted brickwork	0.01 - 0.02
3 mm plywood panel	0.01 - 0.02
6 mm cork sheet	0.1 - 0.2
6 mm porous rubber sheet	0.1 - 0.2
12 mm fiberboard on battens	0.3 - 0.4
25 mm wood wool cement on battens	0.6 - 0.07
50 mm slag wool or glass silk	0.8 - 0.9
12 mm acoustic belt	0.5 - 0.5
Hardwood	0.3
25 mm sprayed asbestos	0.6 - 0.7
Persons, each	2.0 - 5.0
Acoustic tiles	0.4 - 0.8

6.4.2 Calculating Total Room Sound Absorption

The total sound absorption in a room can be calculated by the following formula:

$$A = \Sigma S_i \alpha_i = S_1 \alpha_1 + S_2 \alpha_2 + .. + S_n \alpha_n$$

where:

A = the absorption of the room (m^2, or "sabin")

S_i = area of each surface (m^2)

A_i = absorption coefficient of the material for each surface

n = total number of surfaces in the room

Sound absorption is usually expressed in units of *sabins,* after the physicist Wallace C. Sabine, who is generally considered to be the father of acoustics theory.

For reference purposes, 1 sabin is the sound absorption equivalent to an open window that is 1 m^2 in area; the larger the calculated absorption for a given room, the more acoustically "dead", or "anechoic", the room will be. Conversely, the smaller the calculated absorption, the more "live" the room will be, and the more it will tend to reverberate sounds that are produced within it.

6.4.3 Calculating Reverberation Time for a Room

Once the total sound absorption has been calculated for a given room, and the volume of the room in m^2 is also known, it is possible to calculate *reverberation time,* defined as the time required for a sound to diminish in intensity by 60dB, by the following formula, developed by Wallace Sabine in 1922:

$$RT_{60} = 0.161\ V / (\Sigma\ S_i\ \alpha_i) \text{ seconds}$$

The reason RT_{60} is calculated on the basis of a sound decaying by 60dB is because the loudest sound that would typically occur within a room is approximately 100dB, and the typical masking threshold (the level below which quieter sounds cannot be perceived) is approximately 40dB. Hence, the 100dB initial sound dying down to below 40dB would require a reduction of 60dB.

Typical reverberation times for rooms of different types are provided in Table 5.

Table 5: Typical Room Reverberation Times (seconds)

Room Characteristics	**Very Soft**	**Soft**	**Normal**	**Hard**	**Very Hard**
Reverberation Time – RT_{60}	$0.2 < RT_{60} < 0.25$	$0.4 < RT_{60} < 0.5$	$0.9 < RT_{60} < 1.1$	$1.8 < RT_{60} < 2.2$	$2.5 < RT_{60} < 4.5$
Typical Room	Radio and TV studio	Restaurant Theater Lecture hall	Office Library Home Apartment	Hospital Large Church	Gymnasium Factory Jail Cell

Rooms that have long RT_{60} reverberation times will generally provide poor speech intelligibility when listening at a distance, due to strong vowel sounds masking subsequent consonant sounds and the "Haas Effect", which is the brain's perception of the direct and echo paths as separate sounds. To correct such severe reverberation in recordings, large adaptive filters are required to model the acoustic environment and cancel its ultimate effect on the audio.

6.5 Speech Intelligibility in the Acoustic Environment

6.5.1 A-Weighted Signal-to-Noise Ratio (SNA)

Assuming that reverberation, muffling, and other acoustic and/or equipment effects are not significant, a very simple method of assessing speech intelligibility is to use the A-weighted signal-to-noise ratio (SNA), which is calculated as the difference between the A-weighted long-term average speech level and the A-weighted long-term average level of background noise, measured over any particular time, or:

$$SNA = L_{SA} - L_{NA}$$

Experimental measurements of speech intelligibility using this method suggest that an SNA value of at least +15dB will provide complete intelligibility; higher SNA values generally contribute no improvement in intelligibility, but do offer improved listenability and reduced "fatigue factor".

For SNA values of less than +15dB, intelligibility will be compromised, but assessment of precisely how much requires a slightly more sophisticated approach as described in the next section.

A simple method of assessing SNA for recorded speech is to observe the bargraph level difference between speech and non-speech sections of the audio.

6.5.2 Articulation Index (AI)

Articulation Index (AI) is calculated as a linear value in the range of 0.0 (completely unintelligible) to 1.0 (completely intelligible) based on the calculations of SNA in five separate octave bands with center frequencies of 250, 500, 1000, 2000, and 4000 Hz.

Calculation of the AI consists of four basic steps:

1. Measure the effective SNA for each octave band
2. "Clip" the SNA for each band to be no more than +18dB and no less than -12dB
3. Add 12dB to each clipped SNA
4. Multiply each net result by the correct empirically-determined weighting factor from the table below
5. Sum all five weighted values and divide by 30dB

Thus the articulation index can be calculated from the formula:

$$AI = \frac{G_{[i]}}{30dB} \sum_{i=1}^{5} (Lsa - Lna + 12)dB$$

where $G_{[i]}$ represents the weighting factor for each octave band, as listed in Table 6.

The AI measurement provides a slightly more precise assessment of intelligibility based on SNA measurements in multiple bands, but as with SNA assessment alone, acoustic effects such as reverberation are not fully taken into account by the method, except in terms of the octave weightings.

A simple method of determining AI for recorded speech would be to utilize a spectrum analyzer to observe level differences by frequency between speech and non-speech sections of the audio, and then applying the formula.

Table 6: Weighting Factors for Calculating Articulation Index

Frequency (Hz)	Weighing Factor (G)
250	0.072
500	0.144
1000	0.222
2000	0.327
4000	0.234

6.5.3 Percentage Articulated Loss of Consonants (%Alcons)

Unlike SNA and AI, %Alcons is a measure of intelligibility that takes into account the loss of consonant information in speech due to reverberation in the acoustic environment. This form of measurement was developed in the early 1970s in order to aid in the design of public address systems for auditoriums, churches, classrooms, etc.

%Alcons values less than 15% are considered to be acceptable (less than 10% is considered very good), with values greater than 15% indicating degraded intelligibility; at 100%, for example, all consonant information is considered to be lost, and the speech is completely unintelligible.

Calculating %Alcons is a quite elaborate process, normally requiring several measurements in the actual acoustic environment, using standard test sources and special measurement equipment. Obviously such a precise measurement is not practical when all we have is a recording made in the environment, but a ballpark measurement might be made.

For example, assuming that the desired speech source is omnidirectional (which is a valid assumption from the post processing point of view, where we cannot know which direction the talker is facing), the basic equation is as follows:

$$\%Alcons = \frac{200r^2 RT_{60}^2 N}{V}$$

where:

r = estimated distance between the talker and the microphone
RT_{60} = calculated reverberation time based on the room's characteristics
V = calculated volume of the room in m^3 based on the room's measurements
N = ratio of total energy (including all noises and reverberations) to direct energy

If we can obtain some basic physical measurements from the environment in which the recording was made (in order to calculate RT_{60} as discussed previously), and the distance between the microphone and the talker is known or can be estimated, a rough calculation of %Alcons can be made, *if* a value for the ratio of total energy to direct energy, *N,* can also be estimated.

Though admittedly crude, one possible method is to derive N from an SNA measurement (as discussed previously) using the formula:

$$N = \frac{1}{10^{(SNA/20)}}$$

6.5.4 Other Speech Intelligibility Measurements

Other speech intelligibility measurement methods that have traditionally been used for acoustic design work include the following:

Speech Transmission Index (STI): For this measurement, a test signal with speech-like characteristics is used to excite the environment under test. A complex amplitude-modulation scheme is used to generate the test signal. Special test equipment compares the received signal against the known test signal in order to determine the depth of modulation in a number of frequency bands. Reductions in the modulation depth represent a loss of intelligibility.

Rapid Speech Transmission Index (RaSTI): A simpler measurement than STI, a modulated test signal is used to excite the environment under test. A microphone is positioned at the imaginary listener's location. Special test equipment provides a RaSTI measurement in real-time. RaSTI does take into account the effects of reverberation and background noise, but tests only in two frequency bands, with the assumption that the response of the sound system is more than 100 Hz to 8 kHz or higher with a flat frequency response. Often, this creates substantial error in the measurement.

Speech Intelligibility Index (SII): This latest method utilizes a human ear "audiogram" model, again using a special test signal and a receiver in the environment under test. Measurements are supposed to be more accurate than either STI or RaSTI, but lacks the ability to take into account the effects of reverberation as with %Alcons.

Like the original Articulation Index, all of these methods generate an intelligibility number between 0.0 (completely unintelligible) to 1.0 (completely intelligible). However, because they inherently require access to the actual environment with a specially-generated, known test signal, it is not practical to modify them for use on recorded audio material in the forensic audio laboratory.

6.6 Vehicular Acoustics

Recordings made in automobiles suffer much less from the effects of reverberation than those produced in rooms. The typical automobile environment contains carpet, padding, and upholstery, all of which absorb sounds and reduce reflections. By deliberate design, the automobile is a highly "dead" acoustic environment.

So the principal culprit in forensic audio is additive noises from engine, road, and wind noises. The audio sound system may also produce masking music. Engine noise coupling into the passenger compartment is also a potential noise source.

Due to the absorption loss of the carpet, padding, etc., under which the microphone may be installed for concealment purposes, muffling effects similar to those present with room concealments are often present in vehicular environments.

6.7 Outdoor Acoustics

Recordings made out of doors encounter a variety of noises and acoustic limitations. Environmental noises such as machinery, animals, automobiles, and airplanes often impair intelligibility. Wind is a major problem, not only because it produces random broadband noise, but also because it can literally carry the distant sound away before it has a chance to reach the microphone. Fortunately, reverberations are generally not an issue, unless reflective surfaces such as surrounding buildings are present.

Outdoor recordings are usually made from a distance with directional microphones (see Section 7.4) or using a body-concealed microphone. Directional microphones include parabolic and multi-duct *shotgun* mics and are difficult to conceal. Array microphones, such as the DAC SpotMic may also be used, with the advantages that the concealment of the fixed microphone array is easier and the electronic aiming is less obtrusive. All directional microphones have a narrow acoustic beam, much like a spot light, which tends to reject much of the background noises that are present. However, they do suffer from the limitation that they receive not only the talkers' voices within the acoustic beam, but also other noise sources that are in front of and behind the talkers.

6.8 Telephone Acoustics

Many forensic recordings are made over telephone channels. Unfortunately, both dial-up and mobile telephone channels have less than ideal transmission characteristics.

The audio passed has a typical bandwidth of 300 to 3000 Hz. The frequency transfer function is not flat; certain frequencies within the *passband* may be attenuated 10 to 20 dB compared to others. The broadband noise floor on such lines ranges from 10 to 30 dB below voice levels. Crosstalk, echoes, power line hum, and amplitude distortion are also common maladies associated with telephone speech.

In the case of digital mobile phones, G.172 or other digital compression is often utilized. A form of ADPCM, low-level signal information is sacrificed and distortion is introduced in order to reduce the number of bits that have to be transmitted on the channel.

6.9 Body-Worn Recorder/Transmitter Acoustics

Very commonly, forensic recordings are made using body-worn recording and/or RF transmitter devices. Firstly, the microphones are generally concealed under clothing and against skin, which means that the absorption loss is quite high and muffling of the audio can be a major problem, particularly for desired sounds arriving from behind the person wearing the device.

Secondly, there are often "rubbing" noises that occur whenever the person moves; as these are generally random broadband noises that are transitory in nature, adaptive noise filtering is largely ineffective, but software-based transient noise reduction techniques may help to reduce the adverse effects.

Another major issue is equipment-induced noise. For example, when using built-in microphones on a mechanical device such as a microcassette or minidisc recorder, the motor noise from the machine itself is often audible in the recording.

Also, with analog RF transmitters, there is generally static present in the received audio, and dropouts often occur due to the relatively low transmitter power and the distance to the receiver.

A relatively new phenomenon is that of mobile phone RF interference with body-worn recording devices, particularly when external microphones are used. For example, a Nextel® or GSM mobile phone will couple RF into a microphone cable and induce an intermittent "buzzing" noise that directly interferes with the recorded audio. Although shielded microphone cabling does help mitigate this to some extent, the sheer transmission power of these phones makes complete elimination of the problem nearly impossible.

7. MICROPHONE SYSTEMS

Sound consists of a longitudinal wave which radiates outward from its source through the air. This outwardly moving disturbance consists of compressions and rarefactions; the more intense the sound, the greater the pressure differential between peaks and valleys in this pressure waveform.

Microphones are used to convert this audio sound pressure wave into a small electrical signal that varies in sympathy with the incoming wave. This signal is very small, from microvolts to a few millivolts, and must be amplified before it is recorded.

7.1 Microphone Types

A variety of microphones exist. Several are described below:

Carbon microphones — The carbon microphone was invented by Alexander Graham Bell in 1876. This primitive design is still used in certain communication applications, especially in the mouthpiece of telephone handsets. It is durable, rugged, and resistant to changes in temperature and humidity; however, it suffers from distortion, high frequency hiss, and poor high frequency response. Carbon microphones employ a chamber of carbon granules which are compressed by the microphone's diaphragm in sympathy with sound vibrations. The electrical resistance of the granules varies with pressure and thus allows circuit voltage to vary with sound vibrations.

Crystal and ceramic microphones — These designs utilize the piezoelectricity, or "pressure electricity," produced by pressing a crystal substance. They are characterized by medium frequency response and modest distortion, but they can be made to have good frequency response.

Dynamic microphones — Also called pressure or moving-coil microphones, these are very popular in professional applications due to their frequency response and low distortion. This microphone produces a small voltage by the motion of its coil, attached to the diaphragm, in a magnetic field. It does not require power to operate, though amplification is usually required.

Capacitor microphones — Capacitor (or condenser) microphones produce a varying voltage from the changing capacitance between the diaphragm element and a fixed plate. The electret capacitor microphone is used extensively in law enforcement applications. It has good response in the range of 30 to 18,000 Hz. This device, illustrated in Figure 7-1, places the capacitive element in a voltage divider circuit. A built-in amplifier incorporates a high-input-impedance field-effect transistor (FET). An *electret* microphone is a capacitor microphone with a permanent charge on the head capacitor.

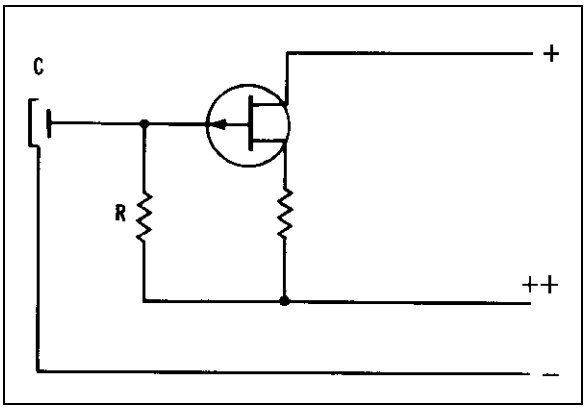


Figure 7-1: Electret Microphone Circuit

Electrets are characterized by their insensitivity to mechanical vibration, flat and extended frequency response, tolerance to loud sound pressures, low noise and distortion, and compact size. Figure 7-2 illustrates the popular Knowles Electronics EK series microphone. This device is used extensively in law enforcement applications.

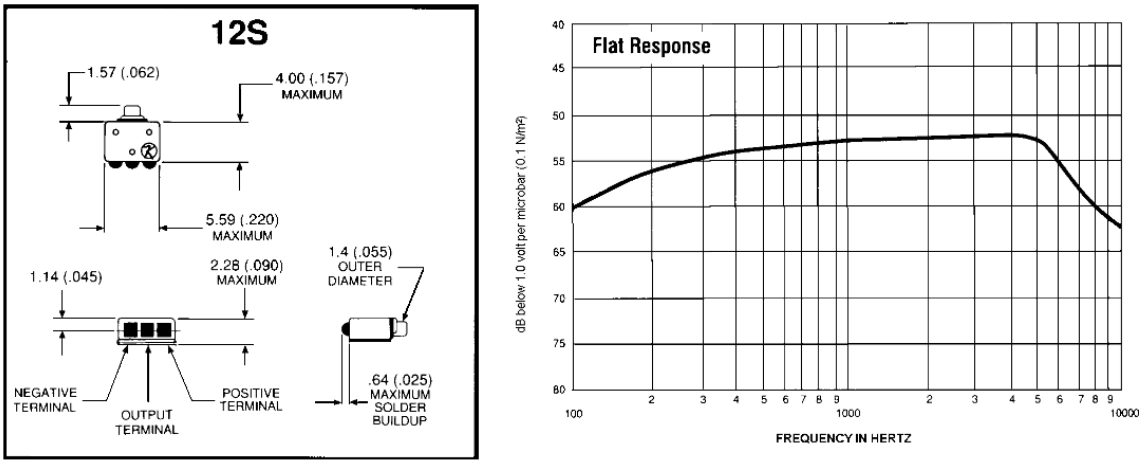


Figure 7-2: Knowles EK Microphone Packaging and Response Curve

Accelerometers — These instruments measure acceleration and thus can be used to measure vibrations. An accelerometer is a transducer that produces an electrical voltage that is directly proportional to the applied force. Fundamentally, it consists of a mass on a spring that is attached to a solid, sound-conducting medium. Sounds in the medium produce vibrations in the mass-spring system. When a constant current source is attached in parallel with the spring, the spring acts as a resistor that varies with movement, and the time-varying voltage that develops across the spring-resistor directly corresponds to the sounds present in the solid medium. Accelerometers can be effective for intercepting sounds that travel through solid objects such as walls, doors, windows, beams, and pipes.

7.2 Room Microphones

When installed in a wall, the location, size, and shape of the wall penetration are critical. These factors largely determine the intelligibility and quality of the audio received. In general, the pin hole should be at mouth level and looking into the room. A larger size pin hole will allow more of the sound pressure to reach the microphone, thus producing a larger signal. The microphone should be installed behind the pin hole and as close as possible to the room side of the wall. Dense acoustic foam should be used to isolate the microphone from the building/wall structure to reduce vibration and building rumble.

7.2.1 Microphone Type, Location, and Transmission

Ideally, the microphone should be installed pointing at the talkers, at mouth level, and away from noise sources. Examples of noise sources are:

1. power outlets and lines
2. fluorescent lights
3. Radio frequency (RF) sources
4. computers
5. machinery
6. heating, cooling, and ventilating equipment
7. radio, television, hi-fi systems

7.2.2 Testing

After each installation or modification, the system should be tested. Listen and be sure that the conversation can be understood with the different noise sources active. If voice intelligibility and audio quality are not satisfactory, try other options. It is best to correct the problem before the recordings are generated.

It is extremely important to utilize high quality microphones and to select a microphone port configuration (snout, top, or front) optimized to the pinhole and acoustic coupling.

7.2.3 Audio Transmission

All audio cables should be shielded, as short as possible, and routed away from noise sources such as:

1. AC power lines
2. AC outlets and transformers
3. electric motors
4. fluorescent lights
5. computers
6. RF transmitters (*e.g.*, cellular phones)

The cable should also be immobilized and located outside of foot traffic patterns.

Several types of communication procedures are commonly used to transmit the microphone audio to its destination. These are summarized below.

- **Telephone** (Dial-up pair) – Use a commercial quality telephone line interface device to isolate the DC voltage potential (and ring voltage) from the connected equipment and impedance match the interface device to the telephone line. This insures that maximum signal is transferred from the telephone line to the audio receiving device. The loading of the interface device interferes with normal telephone line operation.

- **Dry line** or **hardwired** – A pair of wires that runs between the target and monitoring point. These wires may be installed by the user or leased from the telephone company.

- **RF** (Link) – A frequency should be selected that will not interface with other services and is free of interference. UHF transmitters have better building material penetration; therefore, the RF signal usually radiates through building construction better than VHF transmitters.

- Examples of other methods of other audio relay links are **optical** and **millimeter**. They are more expensive and are usually harder to install and set up. These systems require special equipment to achieve optimum performance.

7.2.4 Preamplifiers

A good commercial quality microphone preamplifier should be used. The preamplifier must be able to handle a very quiet to a very loud sound. Sounds intercepted in electronic surveillances vary from a whisper to a door slamming. The ability of the preamplifier to handle varying loudness signals (levels) is called dynamic range and is measured in dB. In an investigative

situation, signal loudness may exceed 85 dBA; therefore, the preamplifier must be able to accept input levels in excess of 100 dBA without distortion.

The frequency response of the preamplifier should be wider than the voice frequency spectrum to avoid this device being the limiting factor in the system. A frequency response of at least 200 to 10,000 hertz is recommended.

7.2.5 AGC/Limiter

If the audio is to be recorded on an analog recorder or relayed over a hardwire, telephone line, dry line, or RF link, the dynamic range of the signal must be compressed to match the dynamic range of the selected transmission/recording medium. Consider the example where the dynamic range of a telephone line is approximately 30 dB and the dynamic range of the audio is approximately 85 dB. A dynamic range of 85 dB cannot be relayed over a telephone line; therefore, a compressor (or AGC device) must be used to reduce the dynamic range to approximately 30 dB. A 3-to-1 compressor will reduce the dynamic range of the 85 dB signal to 28 dB which can be relayed over a telephone line. If the dynamic range is not compressed, then significant signal will be lost and/or distortion will be increased. A compressor helps insure that the loudest and softest signals will be available for review and post processing if necessary.

Kinds of equipment used to reduce the dynamic range of signals are referred to as automatic gain controls, limiters, automatic level controls, and compressor/expanders (companders).

7.2.6 Recorders

Digital recorders are capable of faithfully recording the entire audio signal. If a digital recorder is available, it is the best choice. The best analog recorders are not capable of recording the full audio signal and introduce certain compromises.

The wow and flutter, speed drift, distortion, and amplitude variation of analog recorders reduce voice intelligibility and the effectiveness of post processing. A recorder should be selected that will facilitate subsequent enhancement, if required.

Chapter 8 discusses the characteristics of several different voice recorder systems.

7.2.7 Headphones

Headphones are usually used to monitor live audio and recordings. When a conversation is played into a speaker located in a room, the audio quality is degraded by room acoustics such as reverberations. The audio may be further degraded by extraneous noises occurring in the room.

Headphone monitoring reduces or eliminates these effects. Headphones should be used whenever possible.

7.2.8 Speakers

High quality loudspeakers may be required to play the audio signal for large groups of people. An equalizer may be used to compensate for the room acoustics. The aural quality of a speaker reproducing an audio signal in a room is usually determined by critical listening. In most cases, one or two speakers are recommended, since multiple speakers usually require phase matching and precise adjustment of sound field overlap.

7.3 Vehicle Microphone Systems

Vehicles have acoustic characteristics similar to those of a small, "soft" room. The acoustic echo paths will be short. Vehicles also have a wide variety of interfering noise sources:

1. AM-FM radio
2. Cassette player and/or CD player speakers
3. AC-heating-ventilation vents
4. Open windows
5. Ineffective door and window seals
6. Turn signal indicator noise
7. Other dashboard noise status indicators
8. Blower and other electric motors
9. Rain and hail
10. Vibration and rattles
11. Engine noise
12. Exhaust and electrical system
13. RF energy of cellular phones
14. Traffic, sirens, airplanes, and tire/road noise

7.3.1 Microphone Placement

Ideally, the microphone should be placed directly in front of the talkers, at mouth level, and as far away as possible from noise sources. In practice, the microphones are usually installed in the front windshield supports, the headliner molding, or the headliner. If possible, two microphones should be installed. The microphones should be isolated from the metal of the vehicle using dense acoustic foam, or a similar product, and protected from air movement in front of the mic port. Microphones should have RF bypass capacities to prevent RF energy from interfering with the audio electronics.

7.3.2 Recorder Selection

Care should be taken to insure that the recorder does not rattle or vibrate when the vehicle is being driven. The list of preferred recorder types from a performance standpoint is:

- Solid-state: SSABR, FBIRD, etc.
- DAT / NT2
- Nagra, SN, SNST, JBR
- compact cassette
- microcassette

7.3.3 Noises Generated by the Recorder

The recorder end-of-tape (EOT) alarm should be disabled if it exists. The acoustic noise generated when the recorder starts, records, stops, pauses, and reaches EOT should be checked by listening in a very quiet location. The recorder should not generate any noticeable acoustic sounds. A remote off-on switch may extend record time but could cause evidentiary problems if the recording is used in court.

7.3.4 Enhancement of Recordings Made in Vehicles

A variety of enhancement processes are available for vehicle audio recordings. These include the following.

1. 1CH adaptive deconvolution (filtering) to automatically identify, track, and reduce time-correlated noises.

2. Lowpass, highpass, and bandpass filtering to reduce noise outside (above and below) the voice frequency spectrum.

3. Notch, bandstop, and equalization filtering within the voice frequency spectrum to reduce discrete tones, banded noise, and spectrum anomalies within the voice frequency spectrum.

4. Digital spectral inverse and spectral subtraction filtering to automatically analyze the recording, determine which acoustic sounds are always present, and generate a filter to reduce these sounds. The acoustic sounds associated with the conversation vary and are not normally removed.

7.4 Directional Microphones

7.4.1 Parabolic and Shotgun Microphones

Directional microphones acquire audio at a great distance. Two popular styles are the shotgun style and parabolic microphones, illustrated in Figure 7-3 and Figure 7-4.

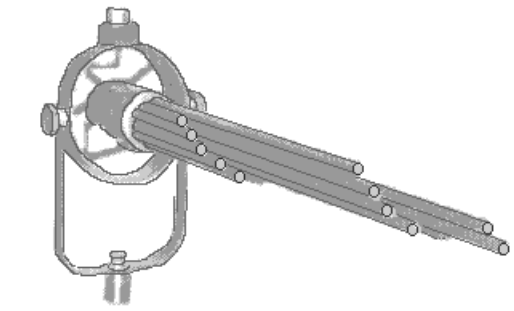


Figure 7-3: Shotgun Directional Microphone

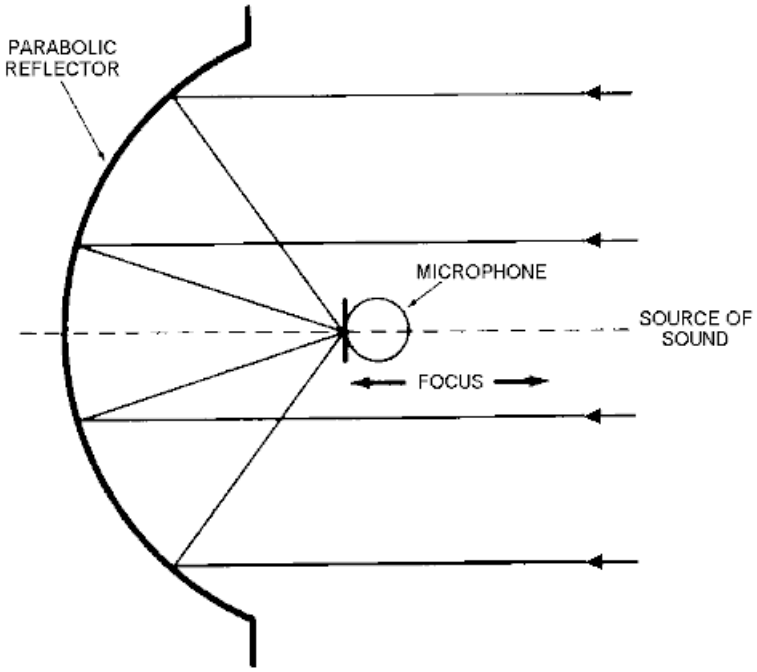


Figure 7-4: Parabolic Directional Microphone

The shotgun microphone, also known as the rifle or tubular mic, consists of a bundle of approximately 20 tubes varying in length from 5 to 150 cm, terminated with a conventional electret or dynamic microphone element. Each tube acts as an acoustic bandpass filter with strong spatial directivity. The bundling of tubes results in an overall bandwidth acceptable for audio applications.

Parabolic microphones place a microphone element at the focus of a reflective parabolic dish. They are sometimes called umbrella microphones due to the umbrella shape of the dish. Gain and directivity are achieved by the effective increase in surface area produced by the dish. Low frequency directivity is affected by the dish diameter. In fact, dishes are not considered directional when their diameter is less than the sound wavelength.

As an example, a dish usable down to 250 Hz would require a diameter greater than one wavelength λ, *i.e.*,

$$\lambda = \frac{\text{Speed of sound}}{\text{frequency}}$$

$$\lambda = \frac{1140 \text{ ft/s}}{250 \text{ Hz}} = 4.56 \text{ ft}$$

The dish must be at least 4.56 feet in diameter.

The gain of a parabolic microphone is proportional to the dish size. As a practical matter, a one-meter dish has approximately 10 dB gain, and a two-meter dish has 16 dB gain.

Neither parabolic nor shotgun microphones are easily concealed, and they have limited applicability to law enforcement.

7.4.2 Linear Array Microphones

Linear array microphones are well suited for law enforcement applications, as they produce significant gain and are readily concealed. The microphone system consists of a sequence of equally spaced pin-hole microphones in a straight line.

A linear array microphone is illustrated in Figure 7-5. The sound source is at a distance and produces planar sound waves which strike all microphones head on. The waves at each mic *add in phase* and result in a microphone gain of

$$\text{Gain} = N \qquad \text{(linear scale)}$$
$$\text{Gain} = 3 \cdot k \text{ dB}, \qquad \text{(decibel scale)}$$

where the number of microphones N is

$$N = 2^k$$

or equivalently

$$k = \log_2 N.$$

A 16-element array thus produces a gain of approximately 12 dB. Waves arriving at an angle do not reach each microphone element at the same time and tend to cancel due to phase differences.

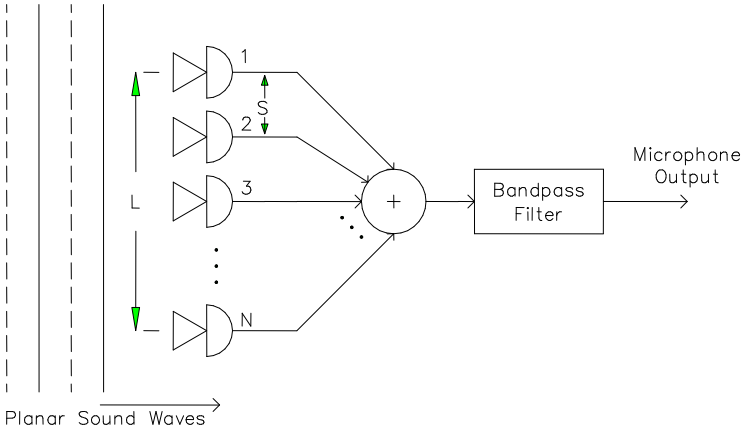


Figure 7-5: Manually-Steerable Linear Array Microphone

The actual pattern of a linear array microphone is symmetrical around the axis of the microphones. The gain pattern is a three-dimensional doughnut with sensitivity forward, backward, up, and down.

The usable, directional bandwidth of a linear array microphone is limited at high frequencies by the element spacing S and at low frequencies by the overall array length L. These frequency limits are specified by

$$\text{frequency} = \frac{c}{l} = \frac{\text{speed of sound}}{\text{length}}$$

As an example, a 16-element array having spacing S = 3 in and length L of 3 × 15 in = 45 in has a useable bandwidth from

$$F_{LOW} = \frac{1140 \text{ ft/s} \times 12 \text{ in/ft}}{45 \text{ in}} = 304 \text{ Hz}$$

to

$$F_{HIGH} = \frac{1140 \text{ ft/s} \times 12 \text{ in/ft}}{3 \text{ in}} = 4560 \text{ Hz}$$

The microphone system will receive audio outside this band but will not have the desired directional properties. To assure directional effectiveness, a bandpass filter (passing 300 to 4500 Hz in the above example) is often placed in the microphone system's output.

Moreover, spatial aliasing can occur at high frequencies unless the relationship between frequency and element spacing is

$$F_{HIGH} < \frac{c}{S \cdot (1 + \sin|\theta_{\max}|)}$$

where $\theta_{\max}$ is the maximum look angle with broadside being θ = 0. Spatial aliasing is comparable to the time aliasing that results from time-sampling at a rate below the Nyquist rate (see Section 10.1.1), and it produces aliased beams known as *grating lobes*,

Linear array microphones may be either manually or electronically steerable. Manual arrays, as shown in Figure 7-5, must be physically rotated to broad-side the sound source of interest. Electronically steerable arrays, such as the DAC SpotMic system, use digital signal processing (DSP) technology to electronically "rotate" the array to the desired direction without physically moving the array. Figure 7-6 illustrates the principle of an electronically steerable array in the form of a delay-and-sum (DS) beamformer.

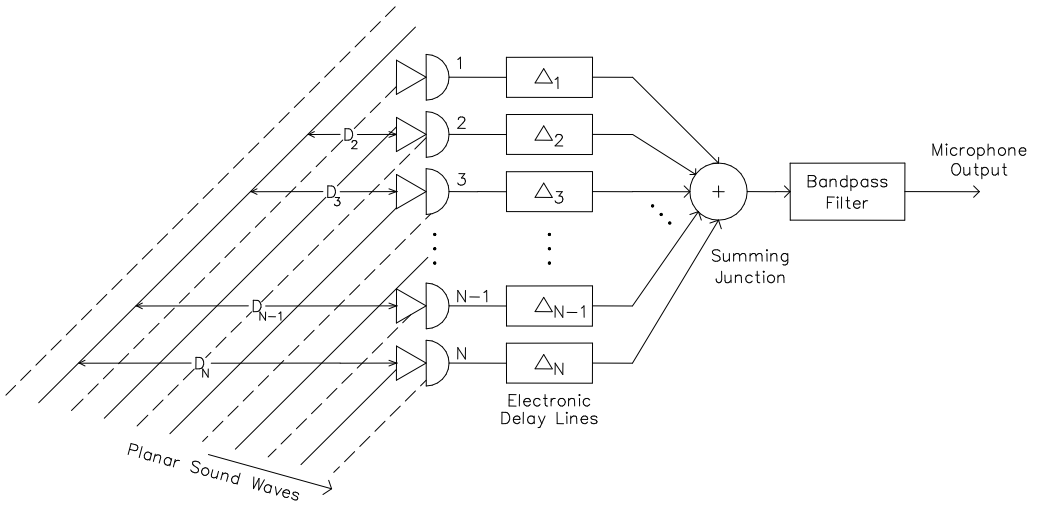


Figure 7-6: Electronically-Steerable Linear Array Microphone (e.g., DAC SpotMic System)

In the figure, delay lines are used to compensate for delays D_2, D_3, . . ., D_{N-1}, and D_N. (Microphone 1 in the above example has a delay of zero.) In so doing, the angular sound wave is in phase at the summing junction. The incoming sound waves can be assumed to be planar if the range from the microphone array to the source is sufficiently large. The minimum range R for safe use of the *far-field* approximation depends on the accuracy needed for the target beamforming pattern, but for many patterns the range should be

$$R \geq \frac{2L^2}{\lambda} = \frac{2L^2 F}{c},$$

where L is the total length of the array, F and λ are the frequency and wavelength of the signal of interest, and c is the speed of sound. For this far-field case, the actual delay for the k^{th} microphone D_k is given by the following equation:

$$D_k = \frac{(k-1)\,S}{\text{speed of sound}} \sin\theta$$

where θ is the angle of the sound source, and S is the individual microphone spacing. If the range is short enough to be *near-field*, then the incoming sound waves must be treated as spherical, and the equation for calculating the delays D_k must consider the range.

In Figure 7-6, the electronic delay lines compensate for the acoustic delays D_2, D_3, etc., which makes the sound wave look broadsided at the summing junction. The delay lines are adjusted as follows:

$$
\begin{aligned}
\Delta_1 &= D_N \\
\Delta_2 &= D_{N-1} \\
&\vdots \\
\Delta_{N-1} &= D_2 \\
\Delta_N &= 0
\end{aligned}
$$

The DS technique outlined here is classical narrowband one-dimensional beamforming. More advanced beamforming techniques include delay-and-filter (DF), two-dimensional, three-dimensional, broadband, statistically optimum, adaptive, beam-shaping, and non-uniform spacing beamforming.

7.5 Laser Listening Devices

Lasers devices can be used to receive audio with a system sometimes called the "laserbug." The laser beam does not need direct access to the targeted area to retrieve the audio, but the laser can be aimed through a window and focused on a hard reflective surface in the room. Acoustic waves in the room cause the window to vibrate slightly, and these vibrations in turn modulate the laser beam. A high-powered optical receiver translates the modulated coherent light into audio. These lasers are usually tuned to the infrared portion of the spectrum ($\lambda \approx 800$ nm) so that they are invisible to the naked eye.

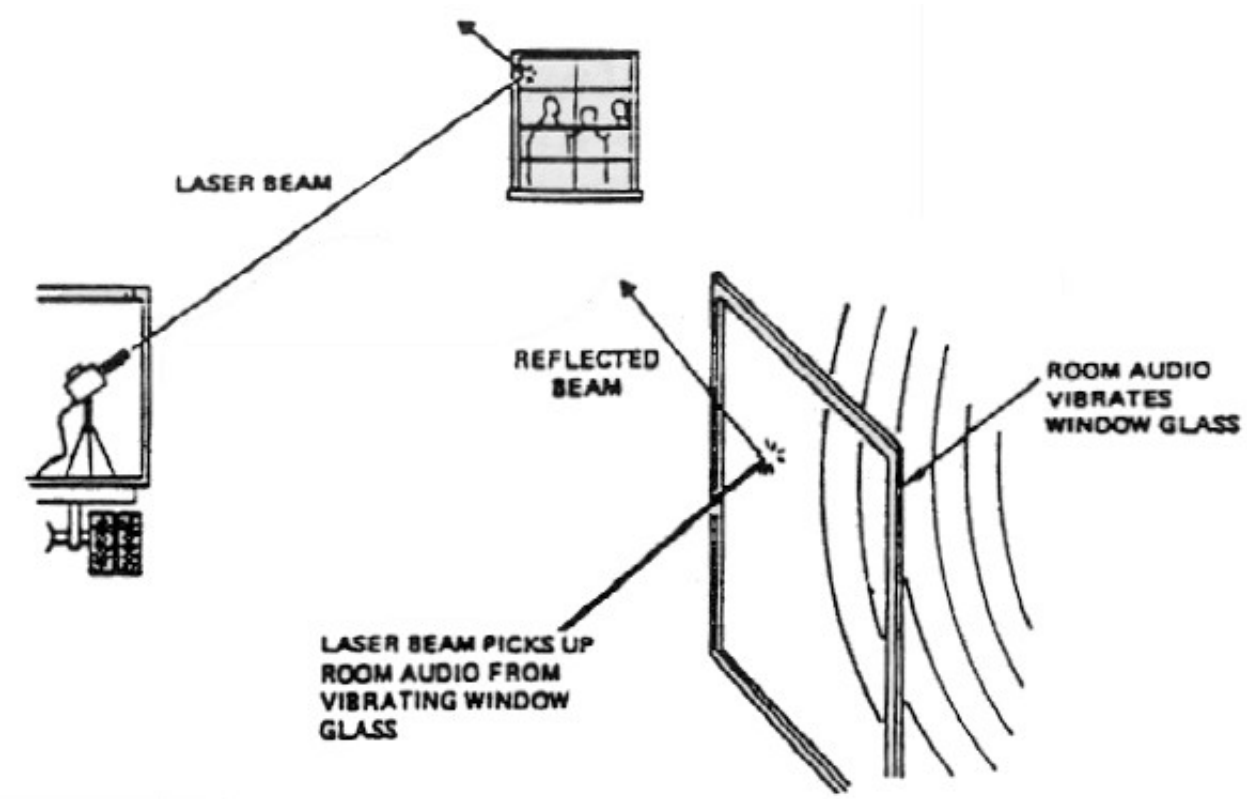


Figure 7-7: Laser Window Pick-Off

Laser listening devices are high-precision instruments, and they are effective only under ideal conditions. They require a steady base (solid surface and tripod) to maintain the correct angles or else they will record noise. Moreover, the effectiveness of these laser devices is limited in daylight and is of short range. Wind, precipitation, and traffic noise can also cause sound interference. Since the device must be aimed directly at an appropriate window, visibility of the instrument is liable to arouse suspicion, though the system may be hidden inside the housing of a camera or telescope.

EXERCISES

1. Of the four types of microphones discussed in Section 7.1, which is most popular for law enforcement applications?

2. What is the wavelength of a 1 kHz tone in air, saltwater, and steel?

3. What is the theoretical gain of a linear microphone array having 64 equally-spaced elements (microphones)?

8. CHARACTERISTICS OF VOICE RECORDERS

Forensic recording relies primarily upon analog instruments. Such instruments have inherent limitations that affect subsequent enhancement and noise cancellation processing of audio. Digital audio tape (DAT) recorders overcome the majority of the technical disadvantages of analog recorders but are substantially more expensive and complex.

8.1 Analog Tape Recorders

8.1.1 Principles of Operation

Direct mode analog tape recorders are popular in law enforcement applications. A functional block diagram of such a recording *system* is given in Figure 8-1 below. Note that this figure illustrates both the record and playback operations; normally only one such operation is carried out at any one time.

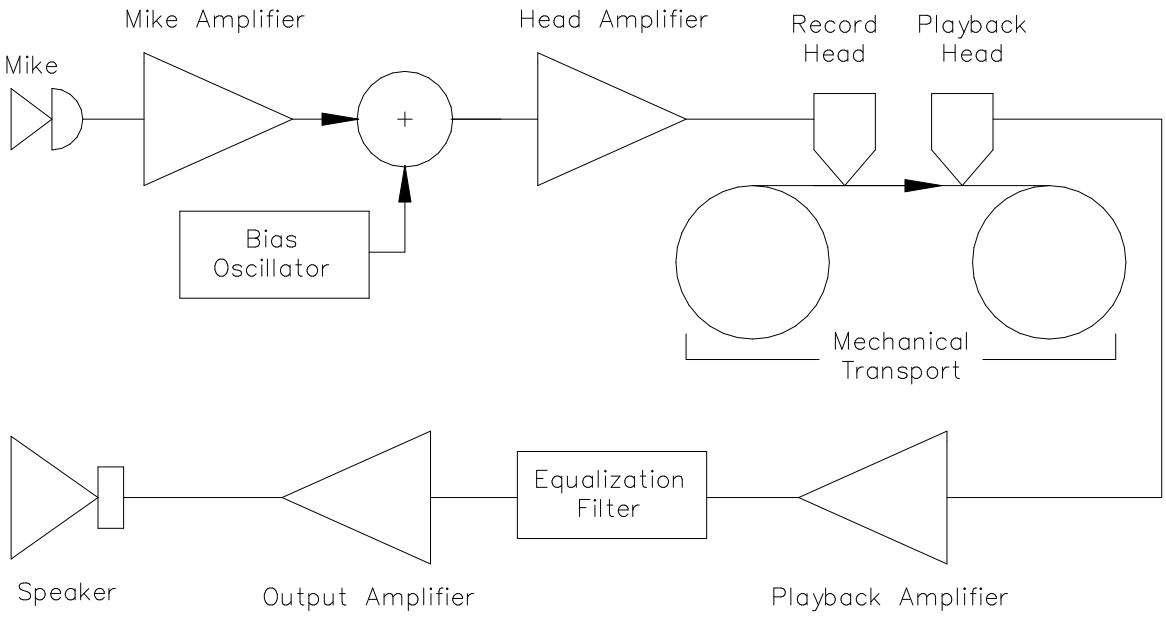


Figure 8-1: Functional Block Diagram of Analog Tape Recorder

The microphone audio is amplified and added to a high frequency bias tone. The sum of these two signals is then imposed on to magnetic tape moving across a record head. During playback, the magnetic field on the tape is induced into the playback head, resulting in a weak voltage corresponding to the original microphone audio. This voltage is then *equalized*. Equalization compensates for the poor low frequency response characteristic of direct playback.

8.1.1.1 Magnetic Tape

Magnetic tape consists of a thin ribbon of polymer plastic material on which a fine magnetizable powder has been deposited using a glue-like binder agent. The plastic material is the base and has no magnetic properties. The binder, likewise, has no magnetic properties. Typically a ferrous oxide material is the magnetic agent, but other materials such as chromium oxide are also used.

The magnetic material should be magnetically *hard*, meaning that it should be able to accept a magnetic orientation from the recorder head and retain that pattern. The material is finely powdered and suspended in the binder, thus allowing each particle to become individually magnetized by the head.

These magnetic particles have an induced magnetic property illustrated by the B-H curve, illustrated below.

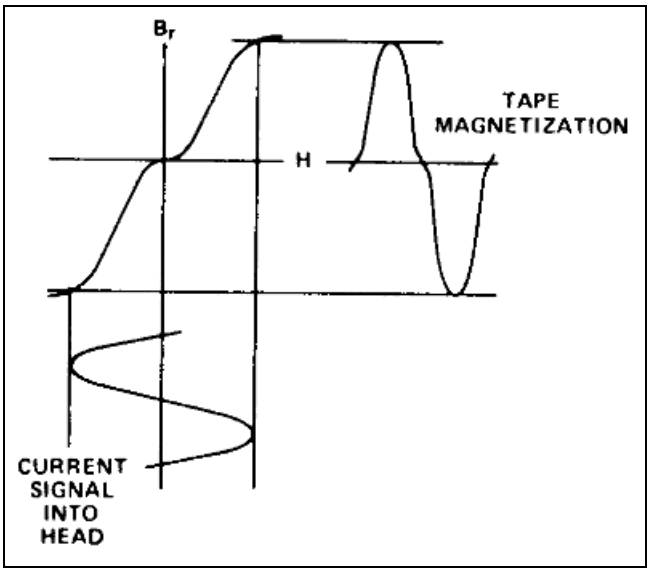


Figure 8-2: Magnetic Tape B-H curve

This curve illustrates that the tape's resulting magnetism, known as flux, is not strictly proportional to the audio applied to the record head. H represents the input drive intensity, and B is the output magnetic flux density. At increasing audio levels (larger values of H), the magnetic material does not yield correspondingly greater level of magnetism (B).

The *linear*, or most straight-line, regions are identified in the figure and are the most desirable ranges into which the tape should be magnetized. If audio were recorded over the entire magnetic range of the tape, shown by the B-H curve, loud (large H) and soft (small H) components of the audio waveform would be greatly distorted.

8.1.1.2 Bias

By adding a bias tone (high-frequency sine wave) at a frequency typically 5× the highest audio frequency being recorded, the audio waveform will be effectively shifted, or *biased*, to occupy the linear (straight-line) region on the B-H curve shown in Figure 8-2. A bias oscillator voltage is added to the audio voltage, and the sum of these two signals is applied to the record head. Since the record head is capable of recording very high frequencies, both the audio and bias signals are recorded on the tape. Unlike the playback head, the record head is not bandwidth limited by head gap spacing.

The playback head is indeed band limited. It, along with bandlimiting electronics, *filter* out the high frequency bias component on playback, leaving only the desired audio. Since the linear region of the tape's magnetic material is used, playback is accomplished with minimal B-H distortion.

The biasing process is illustrated in Figure 8-3:

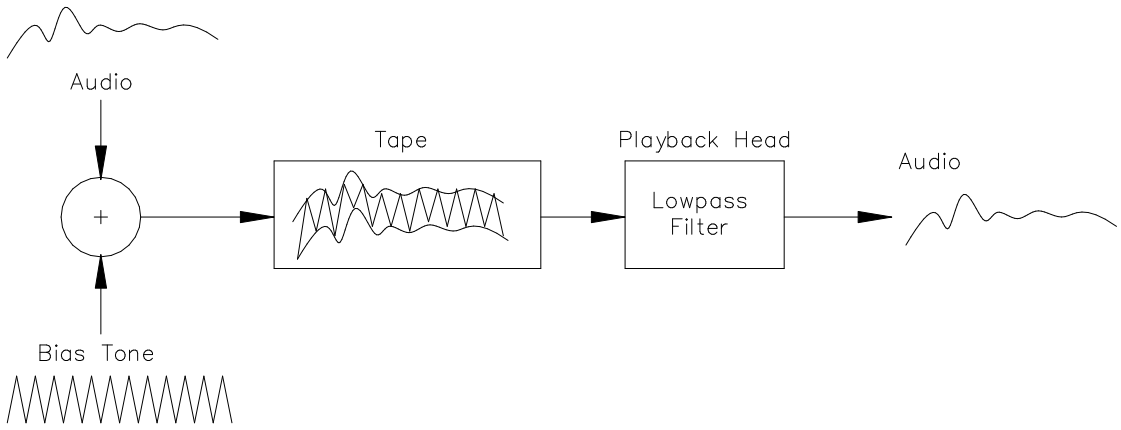


Figure 8-3: Functional Representation of Tape Biasing

8.1.1.3 Tape Heads

In inexpensive tape recorders, the record and playback functions are carried out by the same head mechanism. In general, two separate heads are used. Professional recorders also add a third head for erasure. The playback head is the most critical head since it has bandwidth limitations and nonuniform amplitude transfer characteristics.

For any single frequency, its wavelength λ on the magnetic tape is given by tape speed v and frequency F, *i.e.*,

$$\lambda = v / F$$

Example:
What is the wavelength λ on a magnetic tape of a 1 kHz tone at 1 $^{7}/_{8}$, 3 $^{3}/_{4}$, and 7 $^{1}/_{2}$ ips?

at 1 $^{7}/_{8}$ ips
$\lambda = (1\ ^{7}/_{8}) / 1000 = 0.001875$ inches or 0.048 cm

at 3 $^{3}/_{4}$ ips
$\lambda = 0.00375$ inches or 0.095 cm

at 7 $^{1}/_{2}$ ips
$\lambda = 0.0075$ inches or 0.191 cm

Note that the wavelength λ increases with tape speed.

The playback head gap length specifies the largest magnetic wavelength reproducible by the head. Since faster tape speeds *v* produce longer lengths, faster tape speeds usually result in wider recordable bandwidths.

The head gap actually functions as a lowpass filter as illustrated below in Figure 8-4.

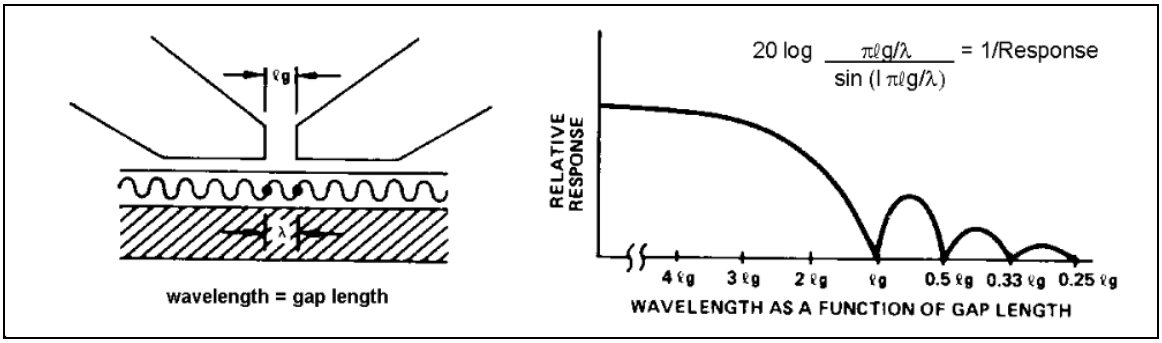


Figure 8-4: Tape Head Gap Function as Low Pass Filter

At lengths substantially longer than the gap length l_g, the audio is not attenuated. At a length of l_g, one complete magnetic wave exists between the gap's two pole pieces and the magnetism is completely cancelled. At even shorter lengths, partial or total cancellation occurs.

8.1.1.4 Equalization

Given two tones, one at 1 kHz and one at 2 kHz, recorded on a tape at equal level, the playback head will reproduce the 2 kHz tone twice as loudly as the 1 kHz tone. In fact, the transfer characteristics of the playback mechanism is characterized by *higher-frequency, higher amplitude*.

The effect can be understood by considering an electric generator. If the shaft is rotated slowly (low frequency), the output voltage is low. When rotated more rapidly (higher frequency) the output voltage increases proportionally. The magnets in the generator are the same; only the rate at which the magnetic field is changing has increased.

Because of this effect, a network which attenuates high frequencies and amplifies low frequencies is required. Figure 8-5 illustrates this *equalization* process

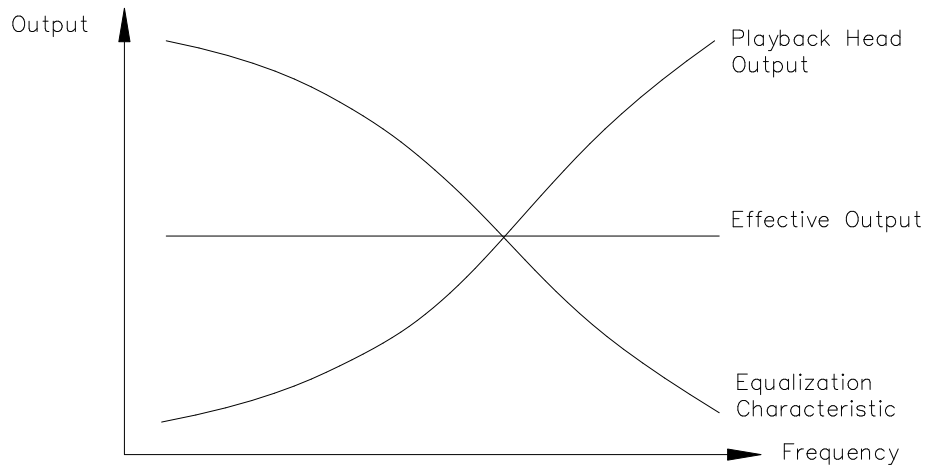


Figure 8-5: Tape Equalization Effects

8.1.1.5 Mechanical Transport

Tape is moved across the heads using a system of reel and capstan drive motor(s), guides, and tension devices. The goal is to pass the tape over the head at a uniform rate of speed. If the record or playback speed is not accurate, the resulting audio will be frequency shifted. A common means of correcting speed inaccuracies on a recording is to look for *characteristic* frequency components using a spectrum analyzer and to adjust the playback speed for correct reproduction of these component frequencies. Examples of characteristic components include

- 50 or 60 Hz power components,
- DTMF (tone dialing) frequencies, and
- known background tone frequencies.

Unfortunately, the tape speed is not precisely uniform. Eccentricities in the capstan mechanism, nonuniform tape drag, and motor speed fluctuations contribute to *wow and flutter*. The effect of these dynamic speed fluctuations is to compress and expand the waveform on the tape. This is a modulation effect and has detrimental effects on reproduced audio quality and the ability to enhance the audio.

8.1.1.6 Increasing Dynamic Range via Companding

Analog tape has a limited dynamic range of 50 to 60 dB. Dynamic range is the difference in the loudest signal recordable (without excessive distortion) to the lowest recordable signal (not masked by the tape's hiss noise floor).

The usable dynamic range may be increased by boosting low level audio before recording and attenuating the same audio on playback. In so doing, the low level audio is recorded at an elevated level above the tape's noise floor. This process is called compression and expansion, or simply *companding*.

Commercial systems such as Dolby and DBX noise reduction employ this technique but are optimized for music reproduction. The system used in Nagra recorders compresses the audio at a 2:1 rate, *i.e.*, each 10 dB increase in volume increases results in a 5 dB increase in record level on the tape. This system works well with voice audio.

Figure 8-6 illustrates a 2:1 compression (100 dB to 50 dB) of the input signal (right side) and a complementary 1:2 expansion (50 dB to 100 dB) of the output signal (left side). Compression and expansion must be precisely complementary for faithful reproduction.

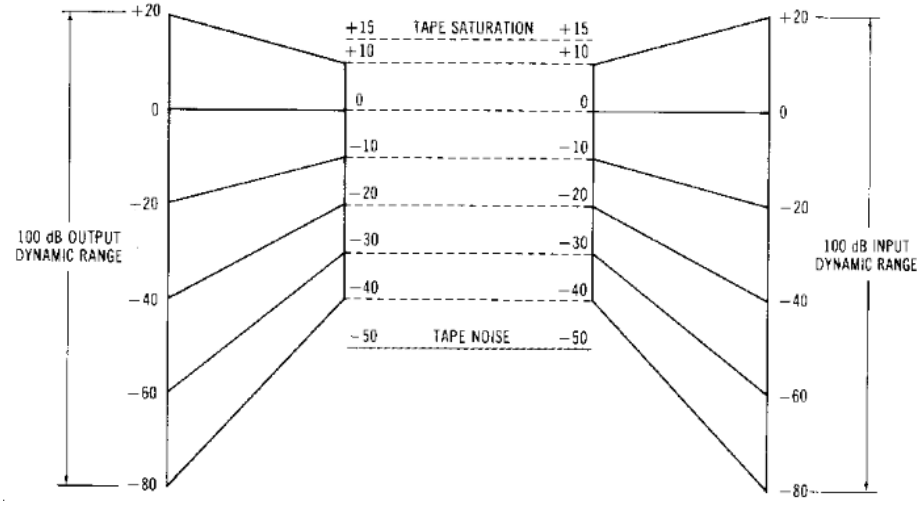


Figure 8-6: Recorder Companding Process

8.1.2 Law Enforcement Analog Recorder Characteristics

The three principal types of analog recorders used in law enforcement are open reel, compact cassette, and microcassette recorders. All three utilize the same *direct recording* principals and differ mainly in the packaging of the recorder and their *media*, *i.e.*, tapes. Tape, consisting of a ferrous oxide material coated on a film base, passes over a magnetic head and is magnetized with a pattern that corresponds to the audio form. Analog recorders are characterized by the speed of tape movement, equalization, track format, and bias. Open reel recorders range in speed from $^{15}/_{16}$ to 30 inches per second (ips). Cassette tapes have a standard speed of 1 $^{7}/_{8}$ ips; microcassettes have two standard speeds, 1 $^{7}/_{8}$ and $^{15}/_{16}$ ips.

The highest recordable frequency is determined by two principal factors: tape speed and head gap. The faster the speed and the narrower the gap, the higher the recordable bandwidth. Small compact cassette recorders are capable of bandwidths of 5 kHz to 10 kHz. Microcassette recorders are designed for voice and have bandwidths of 2.5 kHz to 5 kHz, depending upon the model and speed. Some modified microcassette recorders have speeds of $^{15}/_{32}$ ips for long record times. Because of their slow speed, they have further reduced bandwidths.

A characteristic of analog recording is that not all frequencies are recorded and played back at the same intensity at which they were recorded. Low frequencies have less amplitude on playback than higher frequencies. For this reason, an equalization circuit is placed in the playback path to boost low frequencies and attenuate high frequencies. Because these transfer characteristics are affected by tape speed, a different equalization circuit is required for each tape speed selected.

Analog tape recorders have a variety of tape widths and track *formats*. Laboratory open reel machines use $^{1}/_{4}$ inch wide tape and operate at 1 $^{7}/_{8}$, 3 $^{3}/_{4}$, and 7 ips. Because of the inconvenience in using these recorders, compact cassette decks are more commonly used. Three body worn reel-to-reel recorders are produced by Nagra in Switzerland. These precision but expensive recorders include a miniature open reel (SN and SNST models) and a proprietary cassette (JBR). The SN is a mono recorder, and the SNST and JBR are stereo recorders. All three use 0.15 inch wide tape with a speed of $^{15}/_{16}$ ips and achieve bandwidths of 5 kHz or better. The SNST and JBR incorporate compressor-limiter circuits for additional dynamic range; both machines require a special playback unit containing the required expander. The SN has a single track, the SNST has two tracks, and the JBR has two tracks plus a center track used for speed regulation.

Standard compact cassette recorders are used for both laboratory and body applications. These recorders use 0.15" wide tape and have four recorded tracks (two in each direction). Cassette recorders operate at 1 $^{7}/_{8}$ ips. Because of the narrow track separation, these recorders suffer from *crosstalk* limitations of approximately 30 dB.

Crosstalk is the effect of one channel mixing with the second recorded channel due to magnetic coupling in the record head. The crosstalk limitation is observed more when recording two independent signals rather than stereo recording of the same audio.

Microcassettes developed for dictation purposes have been adapted for law enforcement body applications. These recorders also use 0.15" wide tape and have two standard speeds: $1\ ^{7}/_{8}$ and $^{15}/_{16}$ ips. Since these recorders are designed primarily for voice dictation, bandwidth is limited.

Reproduction of audio tapes requires that the playback machine be matched to the original recorder. Often the original recorder is used for playback. Since the head positioning varies between recorders, matching the head *azimuth* (rotation) to the recorded track can improve playback quality. Some recorders have head adjustments to permit aligning the playback head with the recorded tracks.

Analog recorders' playback fidelity is limited by distortion, speed errors, wow/flutter, and noise floor. Other limitations, previously discussed, are bandwidth and crosstalk. Analog recorders usually have *distortion levels* of 1% or worse. This is primarily due to nonlinearities in the B-H magnetization characteristics of tape. The use of AC bias reduces distortion by spreading the recorded audio over the most linear range of the tape.

Speed errors between the tape recorder and tape player shift the pitch of the audio. Many tape recorders have variable pitch controls to correct for such errors. In doing so, the tape player's speed is misadjusted to match that of the original recorder. A convenient way of obtaining a *speed reference* is to observe known frequencies present in the audio with an FFT spectrum analyzer. Often 50 or 60 Hz hum, DTMF tones, or other identifiable tones are present for this purpose.

Wow and flutter are dynamic speed fluctuations in an analog recording. These effects are due to the tape's passing over the record/playback head at a variable speed. Motor speed regulation, varying tape tension, and irregularities in pinch and backup roller shape all contribute. Wow is low frequency (a few Hz), while flutter is high frequency (up to hundreds of Hz) speed variations. Wow and flutter is contributed during both the recording and the playback process. The overall effect of the fluctuations is to produce an undesired modulation effect. Substantial wow and flutter are audible as a vibrating or nervous overtone to the voice and a loss of audio crispness. Even modest levels of wow and flutter impair enhancement and noise cancellation.

The audio voltage waveform is converted directly to a magnetic intensity in the recording head and applied directly to a moving tape. The ferrous particles on the tape act as small *magnets* which are aligned to reflect the pattern and intensity of the head-applied magnetic pattern. At low audio levels (and hence low magnetic intensities) these small magnets are only mildly rearranged and retain much of their original random pattern. This random pattern, when played back, appears as broad based random hiss. The fineness of the ferrous particles and the magnetic properties of the ferrous material determine the level of this hiss noise.

The ratio of the strongest recordable signal to the level of this limiting hiss noise is called the *signal-to-noise ratio* (SNR or S/N). For most direct recordings, this S/N ranges from 40 to 60 dB.

Ferrofluid Examination:

- 0.2 to 1.5 micron particles
- Apply, allow to dry, microscope examine
- Clean with freon-based solvent

Table 7: Playback Characterization

Amplitude losses at 3 kHz for 20-minute azimuth misalignments using different tape formats.[1]

Format	Track	Speed (ips)	Loss (dB)
Standard cassette	1/4	1 7/8	0.7
Standard cassette	1/2	1 7/8	5.0
Standard cassette	1/2	15/16	19.8
Microcassette	1/2	15/16	19.8
Open reel	full	7 1/2	4.7
Open reel	1/2	1 7/8	8.6
Open reel	1/2	15/16	13.3
Open reel	1/4	3 3/4	0.6
Logging reel (1-inch tape)	40	15/32	7.8

[1] Bruce E. Koenig, "Enhancement of Forensic Audio Recordings," *J. Audio Eng. Soc.*, Volume XXXVI, No. 11.

8.2 Digital Tape Recorders

Digital audio tape (DAT) recorders first appeared in the 1970s as expensive professional instruments. Today, machines are available in both the professional and consumer markets. Those used in law enforcement use small cassettes, approximately the size of a standard compact cassette, and record up to two hours of very high quality audio.

A functional block diagram of a digital recording system is given in Figure 8-7.

DATs include both a digital signal processor (DSP) and a data tape recorder. The DSP includes an analog-to-digital (A/D) converter for record-encoding the audio, a digital-to-analog (D/A) converter for playback-decoding, and their associated sampling filters and analog electronics.

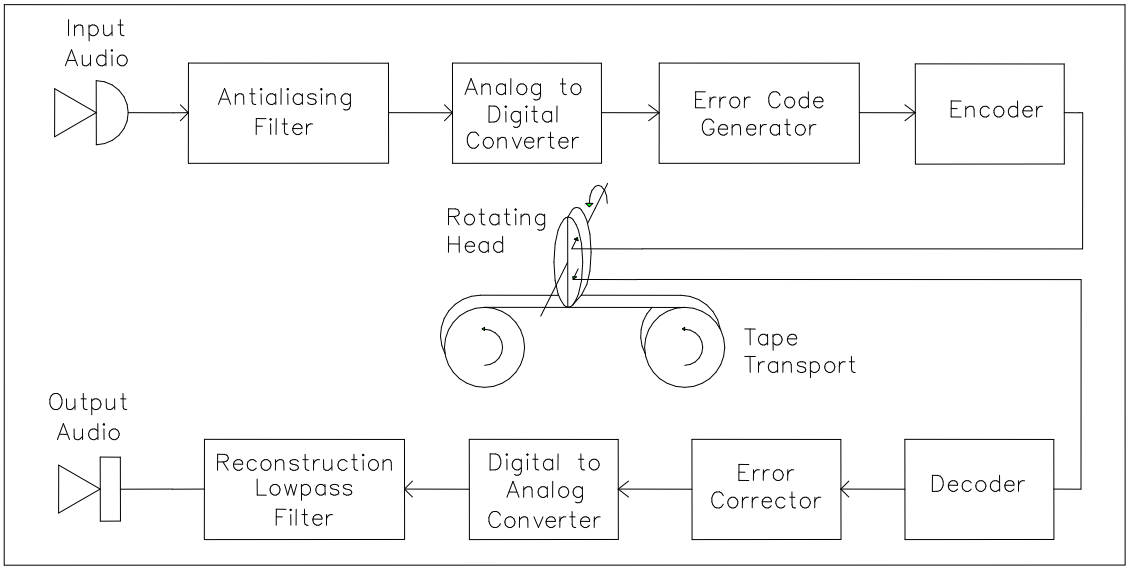


Figure 8-7: Digital Tape Recorder

Input audio is lowpass filtered, typically to 20 kHz, to avoid aliasing distortion (discussed in Section 10.1). The A/D then converts the analog audio into a sequence of numbers. This sampling process makes instantaneous voltage measurements at a rate of either 48,000 or 44,100 samples per second, resulting in a bandwidth in excess of 20 kHz. Each measurement results in a 16-bit sample having a dynamic range approaching 96 dB. Reduced speed DATs sample at 32,000 samples per second and use 14 kHz lowpass filters. They also use 14-bit samples and have a smaller dynamic range. The reduced speed format will record twice as long (up to 4 hours) as the standard format.

The digital samples are stored on tapes as binary (1 or 0) numbers. Since only two levels have to be distinguished, the recorded data can have very poor noise characteristics and still distinguish between 1 and 0. The density of binary numbers on the tape is very high. Errors in reading back numbers can cause objectionable noises in the output audio. For this reason, a system of error-correction coding is used. Here, additional bits are added to assist in both detecting a read-back error and correcting such errors. By adding redundant bits to the data, even tape dropouts causing burst errors can be corrected.

DAT tape is 4 mm wide, and the audio is recorded very densely using a helical scan technique with a rotating drum-type head similar to that of VCRs. DAT tape is available in various lengths. At standard speed up to two hours of continuous record stereo audio may be recorded. At reduced speed the record time is doubled. Even at reduced speed, audio quality is substantially better than analog recorders.

The bits are recorded on tape as a sequence of near-longitudinal stripes (at an angle of approximately 6°). The recording process uses a modulator which is sensitive to abrupt shifts in magnetism. The tape is saturated, as magnetic distortion is not a concern in digital recording. This transition-sensitive modulation (and demodulation on playback) technique is self synchronizing and facilitates the detection of individual bits in each stripe.

On playback, the detected bits are error-detected and error-corrected and are then passed to the digital-to-analog converter for audio reconstruction. The output lowpass filter removes high frequency components inherent in this reconstruction process.

DATs have numerous advantages over analog recorders. Since the A/D process is very carefully controlled, the noise floor is very low. Typical signal-to-noise ratios exceed 90 dB.

No equalization is required, and the DAT has an extremely accurate amplitude response across its bandwidth. DATs are also immune to crosstalk since the stereo channels are separated digitally. Distortion is very well controlled by the analog and digital electronics; the tape's non-linear B-H transfer characteristic does not introduce audio distortion in DATs.

Other DAT advantages include the absence of wow and flutter and extremely accurate speed control. Because the A/D and D/A are crystal oscillator controlled, such degradations are not present.

DAT audio very accurately resembles live audio. These machines are able to reproduce audio so accurately that digital enhancement and noise cancellation procedures cannot distinguish between recorded and live audio.

Sony's NT-2 is a digital micro recorder that records digital stereo onto stamp-size cassette tapes. It samples data at 32 kHz and has a playback frequency range of 10 Hz to 14,500 Hz. The tape speed is approximately 6.35 mm/sec, and its playback dynamic range is better than 80 dB. The tiny cassette tapes can record 90 or 120 minutes, and the seamless auto-reverse feature

automatically changes direction at the end of the tape without interrupting recording or playback. The small size of the recorder and cassette tapes makes this device popular in the law enforcement community. One key concern with the NT-2, however, is that the tapes are extremely fragile. Ideally, once recorded, they should be played back only *once* to provide a direct digital copy onto a standard full-size DAT tape. Repeated playbacks of an NT-2 tape, especially during tape enhancement, may cause the tape to break and data to be lost.

8.3 MiniDisc Recorders

In 1992, Sony unveiled the MiniDisc (MD), and now several companies manufacture MiniDisc recorders and players. The MiniDisc is a digital magneto-optical format that avoids several of the disadvantages of tape formats, such as susceptibility to breakage, limited shelf life, and sound quality deterioration over time. An MD allows for recording of multiple tracks but, unlike a tape device, does not require that tracks be played back sequentially. MDs store data like a hard disk or floppy disk in a computer; tracks can be erased, split, combined, moved, and named. Despite these advantages, however, the lossy compression of MDs makes them unsuitable for recording audio that will require subsequent enhancement in the laboratory.

MiniDiscs are ultra-compact 2 ½" (64 mm) diameter optical discs. Most discs can store up to 74 minutes of digital audio, though discs are also available in 60 and 80 minute formats. Sampling is at a rate of 44.1 kHz with 16-bit precision. The SNR rating of current (2001) MD models ranges from 92 to 112 dB.

MiniDiscs can hold 140 MB of data, which is only about one-fifth as much raw data as a compact disc (CD); however, a special compression technique makes it possible to fit 74 minutes of high-fidelity stereo digital audio on a MiniDisc. The Adaptive Transform Acoustic Coding (ATRAC) system uses psychoacoustic concepts to reduce the amount of data stored without affecting listener perception. ATRAC introduces a small amount of noise at an inaudible level. Early versions of ATRAC introduced more noise, but the latest versions have demonstrated the same quality as DAT in a blind listening test. This data reduction technique does accurately reproduce the sound as perceived by someone listening to the original, but it does not capture all of the sound which could not be originally perceived by a listener. The algorithm discards parts of the original sound that are less perceptible, such as soft sounds or parts of the audio signal which are masked by louder sounds. The difference in quality is barely audible with careful listening, but it negatively impacts audio enhancement when a loud, undesired noise is kept while a soft, masked voice is thrown out. Therefore, a lossy compression technique such as ATRAC *is not adequate for audio enhancement*. Section 8.5.1 further details the concepts and disadvantages of lossy compression.

8.4 Solid-State Recorders

In the late 1990s, several companies, including DAC, began producing solid-state audio recorder/players. Advances in micro-digital technology, coupled with falling prices of Flash memory, have led to a revolution in these portable digital recorders. In a solid-state storage device, there are no moving parts because all components are electronic instead of mechanical. These devices store audio data directly on solid-state computer memory chips called Flash memory. Data stored in Flash memory chips is nonvolatile, which means that it is not lost when power is removed. Figure 8-8 illustrates the functional block diagram of a solid-state audio system. Solid-state recorders can achieve a high quality of recording, but in a forensic environment care must be taken to consider the concepts of original evidence and audio enhancement when selecting a solid-state recorder.

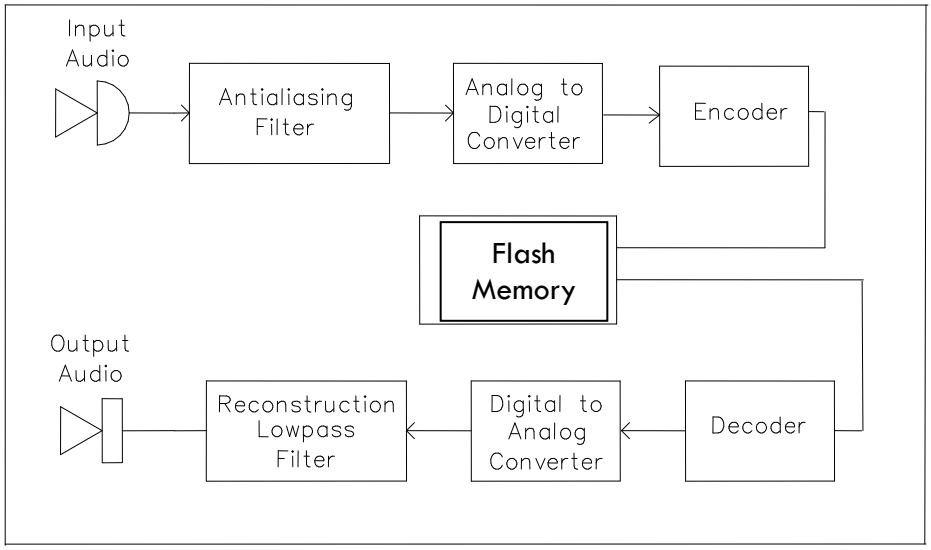


Figure 8-8: Solid State Recorder

These recorders have many attractive qualities, including ease of operation, instant access to any point in the recording, excellent sound quality, and high reliability. They completely avoid all the problems of tape hiss, wow and flutter, head misalignment, and the speed and volume fluctuations that are often associated with the older analog cassette and microcassette machines. Moreover, the packaging of solid-state recorders makes them attractive for law-enforcement and intelligence operations. The devices can be relatively small, and some are smaller than the palm of a hand. Most solid-state recorders are shock resistant and immune to environmental extremes.

Some of these recorders rarely have as much digital storage capacity as would be needed for perfect recording and playback due to size, cost, or battery life constraints. Therefore, they often use imperfect lossy data compression schemes to try to achieve both the necessary record time and, hopefully, an acceptable level of quality for the target application. Section 8.5.1 discusses the weaknesses of lossy data compression for forensic recorders. When audio enhancement may

be necessary for a recording, it is strongly suggested that either a lossless compression scheme (see Section 8.5.2) or no compression be used. Check to see what form of compression is used in a solid-state recorder or whether the compression can be turned off.

When a solid-state recorder is to be used for law enforcement applications, special consideration must be given to the concept of original evidence. Some of these devices use PCMCIA linear flashcards. Though using a PCMCIA card makes it easy to swap the memory in and out, it is impractical in law enforcement due to the need to maintain the original evidence. Because it is a removable medium, the card is considered to be the original evidence. Thus, the flashcards must be stored away rather than reused, and the media is prohibitively expensive. Instead, when the original evidence must be carefully stored, special recorders with non-removable media should be used. The memory chips on these recorders can not be removed, and U.S. courts have ruled that the original evidence is the first copy of the data from the device onto a removable medium. Several solid-state recorders (including the DAC SSABR unit) are designed to support this concept of original evidence for law enforcement.

Solid-state audio recorders are able to achieve a high quality of recording. Though available quality varies by manufacturer, some can perform up to 24-bit 96 kHz sampling in stereo. There is generally a trade-off between quality of audio and the length of recording, and many recorders allow the user to control the quality level desired.

Another feature of solid-state recorders is control of recording timing. In addition to standard record buttons and remote record control, many systems allow for pre-programming of record operations. The user can connect the device to a computer in order to program the recorder to automatically activate later at specified times.

8.5 Digital Audio Compression

Many digital audio storage systems do not have as much digital storage capacity as would be needed for perfect recording and playback due to size, cost, or battery life constraints. In order to store more audio data in less space, they often employ data compression schemes. These compression algorithms are able to reduce the necessary memory storage, but they vary in level of quality. While several systems are able to achieve high quality, not all of them are adequate for forensics applications.

Uncompressed digital audio is often stored in linear pulse code modulation (PCM) format. The two general types of compression algorithms are *lossy* schemes, which compress the data more at the cost of losing some information, and *lossless* schemes, which compress the data less but do not lose any data. Uncompressed data has a compression ratio of 1:1, while a higher ratio indicates greater compression. Lossless algorithms rarely obtain a compression ratio of more than 3:1, while lossy algorithms achieve compression ratios of 4:1 up to 12:1 and higher.

8.5.1 Lossy Compression Schemes

Lossy compression reduces the number of bits that the recorder needs to store by continuously analyzing the original uncompressed digital audio bitstream and discarding as many bits as are necessary to achieve the desired compression ratio. Only the remaining bits are actually stored by the recorder. When the recording is played back, the complementary decompression algorithm attempts to reconstruct the original uncompressed bitstream from that subset of the original bits.

Unfortunately, lossy algorithms fail to reconstruct the original bitstream perfectly; this is precisely why they are called "lossy." Many subtle details of the original uncompressed bitstream will be missing or at least heavily distorted. Thus, substantial audio information that was present in the original uncompressed data stream may be gone or at least degraded, especially low-level signals that may be of great forensic interest.

There are three basic flavors of lossy compression in widespread use today. Of these three, *perceptual encoding* is by far the most common because of the high compression ratios that can be achieved (5:1 or higher). This form of compression is what the newer MP3 and MiniDisc digital recorders use. A diagram that shows how this process works is provided in Figure 8-9.

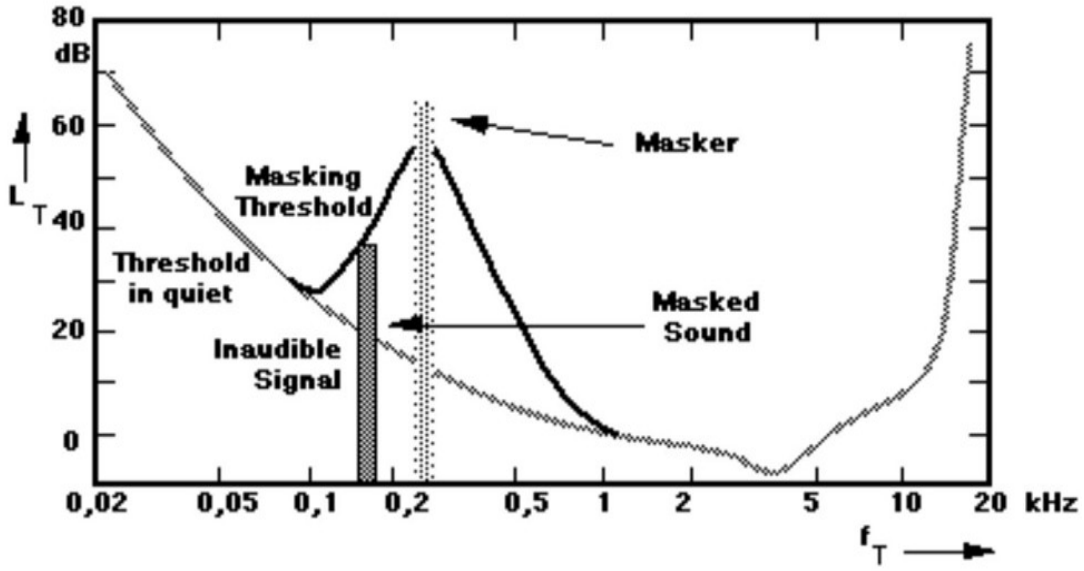


Figure 8-9: Perceptual Encoding

With perceptual encoding, the incoming audio is analyzed using a mathematical model of the human ear. This model assumes that in the presence of loud sounds at particular frequencies, lower level sounds at other adjacent frequencies will not be audible and can therefore be ignored and not recorded. Thus, this method can achieve very high compression ratios (often 20:1) with little perceived loss of original audio quality. This scheme is perfect for re-recording CDs, which

are generally excellent recordings to start with. Remember, this is precisely what MP3 and MiniDisc systems are designed for, and they do it very well.

The problem, though, comes when one attempts to enhance the audio recorded by such a device, to try to improve upon the original. For example, suppose a recording was made in a nightclub next to a powerful loudspeaker, and the conversation of interest is either barely audible or completely masked by the loud music. If the recording was made by a good stereo digital recorder that does not use compression (such as a DAT or CD recorder), we would stand a very good chance of being able to reduce the music and reveal that conversation back in the laboratory using standard two-channel noise filtering (see Chapter 11 and Section 12.3). On the other hand, if the recording was made on a MiniDisc machine, chances are that once the music is taken away by the processing, there will be little or no speech information left to be revealed, as the compression algorithm on the recorder would have thrown all or most of it away when the recording was made, wrongly assuming that no one would ever want to hear it. ***This is the big danger with MiniDiscs and similar lossy compression devices.***

Other lossy compression schemes, which are not quite as problematic as perceptual encoding, include predictive encoding and non-linear sampling. The best example of predictive encoding is adaptive differential pulse code modulation (ADPCM), which is used on most digital mobile phones and many portable solid-state dictation machines. It is also used on the ADS EAGLE and FBIRD recorders; excellent post processing results can be obtained with these recorders when the compression is turned off, but results are compromised when the compression is turned on.

The best examples of non-linear sampling are mu–law (μ–law) and A-law companding. Both coders use 8 bits to represent each sample, and they have an effective SNR of about 35 dB. These schemes are speech-coding standards used by the International Telecommunications Union (ITU) for network telephony; mu-law is used in North America and Japan, while the rest of the world uses A-law. A high-performance 12-bit version of mu-law is used on the Sony NT-2 miniature digital recorder; excellent post processing results can be obtained with this recorder, despite the slight loss in quality relative to normal DAT machines.

8.5.2 Lossless Compression Schemes

To make digital audio recordings that have no loss or degradation of the signal, all the information in the original bitstream must be completely recorded. This can be achieved either by using no compression at all, which is exactly what is done on DAT and CD recorders, or by using lossless compression. Lossless audio compression is a data reduction process in which a digital signal processor (DSP) within the recorder is able to reduce the bits that need to be stored, thus extending the record time, without sacrificing any of the information in the original digital audio stream.

From a forensic audio standpoint, a digital recorder that employs lossless compression is ideal because the storage requirements can be minimized while retaining complete capability to

perform subsequent post-processing and analysis of the recorded audio. Lossless compression always produces exactly the same final product as when no compression is used.

A good example of a lossless compression system is PKZIP, that ubiquitous file compression program that is commonly used to crunch computer program and data files to the minimum possible size before being transferred via modem and/or Internet connection. PKZIP has to be lossless because losing even a single bit of information in a computer program can keep the entire program from running. PKZIP works by intelligently determining which bits in the original file are redundant using a very carefully crafted, patented algorithm. Once these redundant bits are removed, file sizes can be reduced by 50% or more, yielding substantial reductions in both required disk space and transmission time.

However, unlike the perceptual compression used on MiniDisc recorder, which has a guaranteed compression ratio of 5:1 at all times regardless of the nature of the audio data, PKZIP provides a variable compression ratio. In other words, it crunches the bits as much as it can wherever it can within the file, but any bits it cannot crunch without losing information are left intact. In many cases, PKZIP achieves no compression at all, and the output file is no smaller (or sometimes even slightly larger) than the original. Though it works quite well on pure text (ASCII) files, PKZIP is particularly ineffective at reducing the size of digitized audio files. In comparison, the MP3 format is able to compress WAV files and CD audio very effectively, though it does sacrifice low-level information since it uses a perceptual encoding algorithm (see Section 8.5.1).

Several algorithms give the ideal lossless compression that PKZIP provides yet are also effective at compressing audio data. These can be classified as predictive models and transform coding. Figure 8-10 shows how a lossless predictive coder works. Based on the original audio samples, the encoder selects an equation to calculate the predicted waveform that most closely matches the actual waveform. The predictive equation and difference information are then stored in much reduced space. When the audio data is uncompressed, the algorithm determines what equation was used to compress the data, recalculates the equation, and then applies the stored difference values to reproduce the original, uncompressed audio data. Lossy versions of predictive coding (see Section 8.5.1) will throw away some of the difference information, but if all of the difference information is saved, then no information is lost in this process and the original audio can be reconstructed perfectly.

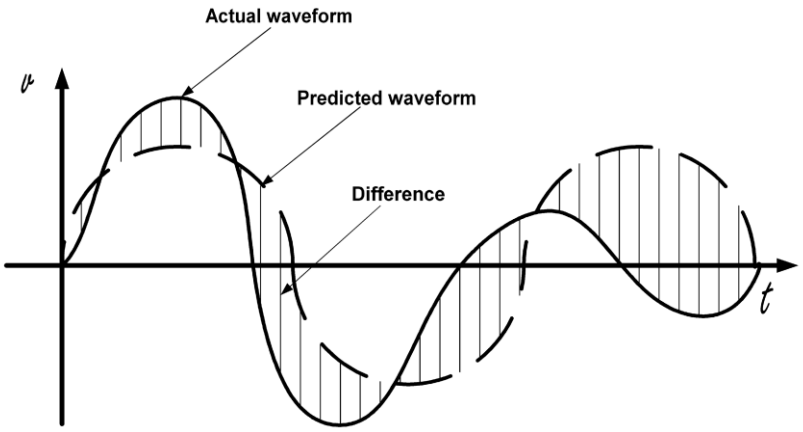


Figure 8-10: Lossless Predictive Coding

As an example, DAC uses a proprietary implementation of the polynomial predictive decorrelation technique, called DACPak, in the SSABR solid-state audio recorder. Unlike PKZIP, DACPak does not see digital audio data as being random, but it instead takes advantage of the special nature of audio. DACPak is able to consider audio data in terms of an equation, as opposed to a set of unrelated numbers. To yield an effective compression ratio, DACPak uses a predictive coding algorithm as outlined above. For stereo audio, DACPak also takes advantage of the fact that there is generally relatively little difference between the left and right channels unless the microphones are spaced widely or completely independent signals are input to the two channels. Thus, compressed stereo audio can often be stored in less than twice the space required by compressed monaural audio.

DACPak typically yields a compression ratio of 2:1 (a 50% bit reduction) on real-world monaural audio. The actual ratio obtained varies, depending upon how closely the selected equation models the actual audio. For example, if the audio signal is a sine wave, the compression ratio will be very high (perhaps as high as 15:1), because such a signal is very predictable, and the equation will produce values that closely match the audio. However, if the signal is random, such as white Gaussian noise, then the compression ratio will be very close to 1:1 (no compression), because that type of signal is very unpredictable. Most real world audio falls somewhere between these two types of signals, so there will generally be some degree of predictability that DACPak can exploit and obtain good compression results.

Table 8: Forensic Voice Recorders

TYPE	ADVANTAGES	DISADVANTAGES
Digital Audio Tape (DAT, NT-2)	Superb quality	Cost, sensitive to vibration and humidity
MiniDisc (MD)	Very good quality	Cost, lossy compression
¼" open reel	Very good quality	Cost, size
Standard cassette	Good quality for the money Easy to load & operate	Most too large to conceal Tape limits to 60 minutes
Micro Cassette	Low cost Easy to conceal	Limited audio quality
Nagras	Very good quality Easy to conceal	Cost
Solid state (SSABR, FBIRD)	Very good quality High reliability	Cost, different concept of "original recording"

EXERCISES

1. What is the wavelength of a 4 kHz tone on a tape recorded at $^{15}/_{16}$, 1 $^{7}/_{8}$, and 3 ¾ ips? If the head gap is 470 millionths of an inch, will the tone play back at all three speeds?

2. What is the purpose of a bias oscillator in an analog tape recorder? Why is the bias signal not present on playback?

3. Incorrect azimuth alignment affects which frequencies most: high or low?

4. List the factors which adversely affect fidelity of an analog recorder.

5. What is the most important maintenance procedure for an analog tape recorder?

9. CLASSICAL SIGNAL PROCESSING

When audio recordings and live signals are contaminated by noise, signal processing is often called upon to reduce the level of the noise. Signal processing generally falls into two categories: *analog* and *digital*.

Analog signal processing consists of analog filters, spectrum equalizers (parametric and graphic equalizers), and analog signal level conditioners (automatic gain control, compressor, and limiter). These instruments rely upon analog circuit components (resistors, capacitors, inductors operational amplifiers, etc.) to improve the audio signal's characteristics. Some analog processors have digital controls and displays but remain, nevertheless, analog signal processors.

Analog instruments have the advantage of lower cost and, in certain situations, are easier to operate than digital signal processors. The digital instruments, however, often provide more precise and robust solutions to audio noise problems.

9.1 Analog versus Digital Signal Processing

Until the late 1970s, digital signal processors were considered far too costly and difficult to implement to solve forensic audio problems. As with the entire electronics industry, digital audio has greatly benefited from the astronomical advances in solid state and microprocessor technology. Such breakthroughs have not only driven the cost of technology downward but have also greatly improved performance.

Audio processors improve signal quality by applying specialized mathematical processes to the signal. One process might attenuate high frequency energy, *i.e.*, lowpass filter, while another might elevate low-level audio, *i.e.*, automatic gain control. Each of these processes is based upon a specific set of mathematical equations developed, in some cases, over 100 year ago. Analog processes attempt to implement these mathematical equations using realizable electronic components. Such implementations are approximations due to the limitations of the components.

Consider a 1000 Ω precision resistor which is used to implement a low-pass filter: the resistor has a tolerance (accuracy) of ±1 percent, a temperature coefficient (the resistance varies with temperature) of 50 ppm, and is subject to effects of aging. All three factors contribute to the inaccuracy of implementation.

Figure 9-1 illustrates a typical analog lowpass filter used in audio applications. The resistor and capacitor values (Rs and Cs) determine that filter's characteristics (cutoff frequency, stopband attenuation, etc.).

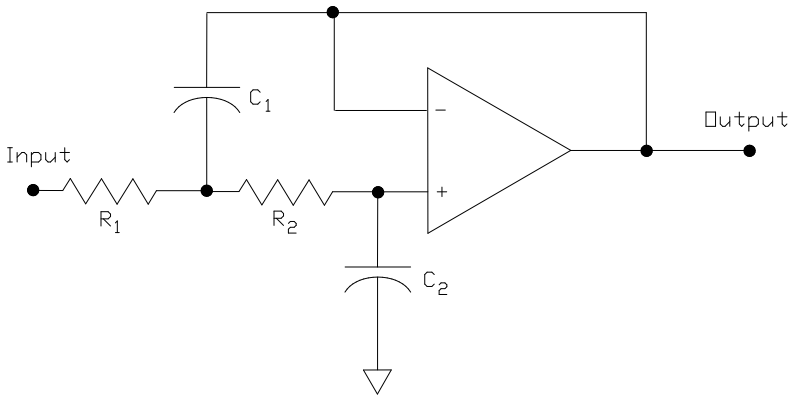


Figure 9-1: Analog Lowpass Filter

Suppose that the analog lowpass filter's cutoff frequency needs to be changed from 1000 Hz to 1200 Hz; that resistor, along with a number of other components, will now have to be changed. Flexibility can often be very difficult to implement in analog circuits. Should this same filter be changed from lowpass to highpass, the resistors and capacitors are interchanged. See Figure 9-2.

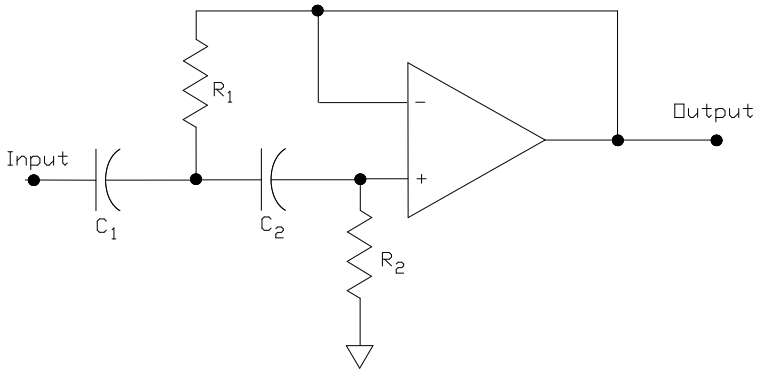


Figure 9-2: Analog Highpass Filter

Consider a digital signal processor (DSP), illustrated in Figure 9-3. In the figure, audio is input into and output from the DSP in *analog* form. Analog signals are continuously varying voltages. These are produced, for example, from microphones and tape recorders. A mic converts the air pressure fluctuations of sound into electrical voltage variations; a tape recorder similarly converts the changing magnetic field of a cassette tape passing over a playback head into voltage fluctuations. Since a digital computer can deal only with discrete numbers, the analog input signal must be converted to a numeric, or *digital*, form.

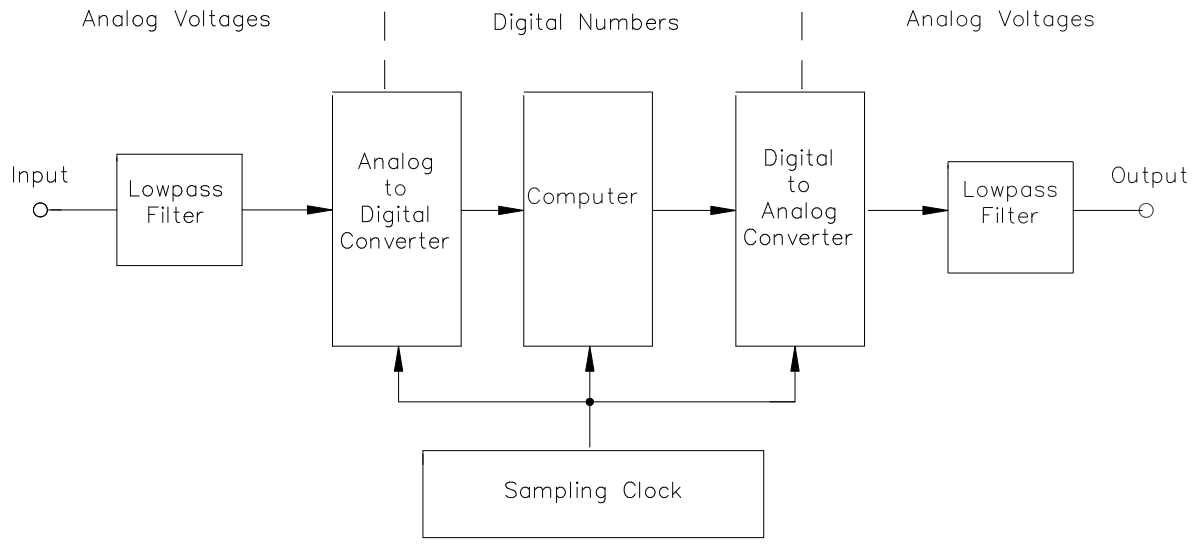


Figure 9-3: Digital Signal Processor (DSP)

A device known as an *Analog-to-Digital Converter* converts the continuous audio waveform into a sequence of digital numbers. Each number represents the voltage of the waveform measured at a specific point in time. This conversion process is known as *sampling*. A *Lowpass Filter* is required on the input to limit high frequency energy. The *Sampling Clock* specifies the rate at which voltage measurements are made. The sample clock frequency must be greater than twice the lowpass filter's cutoff frequency.

Audio samples from the analog-to-digital converter (A/D) are synchronously fed to the digital computer for signal processing. As each A/D sample is received, the digital computer outputs a processed sample to the *Digital-to-Analog-Converter* (D/A). The sampling clock synchronizes this *real-time* digital signal processor.

The D/A reconverts the computer's digital output signal to an analog signal. Inherent in this D/A process is undesired high frequency noise. This noise is removed with the output *Lowpass Filter*. The resulting audio signal may then be passed to headphones, a recorder, or a loudspeaker.

The *Digital Computer* block in Figure 9-3 in many ways resembles a personal computer (PC). It takes in data, processes it according to a specific program, and outputs the results. In this case, the audio samples are the input data, the signal processing procedure is the program, and the improved output audio samples are the results. The program consists of a sequence of mathematical expressions and decisions; as in a PC, this may be easily replaced with another program, or the program characteristics may be easily modified. The 1000 Hz lowpass filter may be changed to a 1200 Hz lowpass by replacing a few key control numbers known as *filter coefficients*. Alternately, the lowpass filter program may be replaced with an automatic gain control program. The personality of the DSP is thus determined by the *software*, *i.e.*, its controlling program. No longer do electronic components need be replaced to change processing

characteristics. In a DSP, the *hardware* is fixed and the *software* is variable. In an analog signal processor, the hardware must be varied since no software exists.

A second, very important advantage of the DSP is its precision. Analog components have accuracy tolerances of 0.1 to 1 percent and are subject to temperature and aging effects. Software values may be specified to *any required accuracy* and do not vary with age or temperature. As a result, processor values may have tolerances of .01 or even .00001 percent! The software need only use more digits in the control program. Since complex filters require great internal precision, digital signal processors are ideally suited for their implementation.

9.2 Conventional Filters

Conventional filters include highpass, lowpass, bandpass, bandstop, notch, and slot filters along with graphic and parametric spectral equalizers. These filters are available in both analog and digital instruments.

9.2.1 Bandlimited Filters (Highpass, Lowpass, Bandpass, and Bandstop)

Bandlimited, or frequency-selective, filters set the amplitude for certain ranges of frequency. These include highpass, lowpass, bandpass, and bandstop filters. The two most basic bandlimiting filters are:

- Highpass Filter (HPF): Passes energy above its *cutoff frequency* f_c; attenuates energy below that frequency.
- Lowpass Filter (LPF): Opposite of the HPF, *i.e.*, passes energy below the cutoff frequency and attenuates energy above that frequency.

These two filters are spectrally illustrated below.

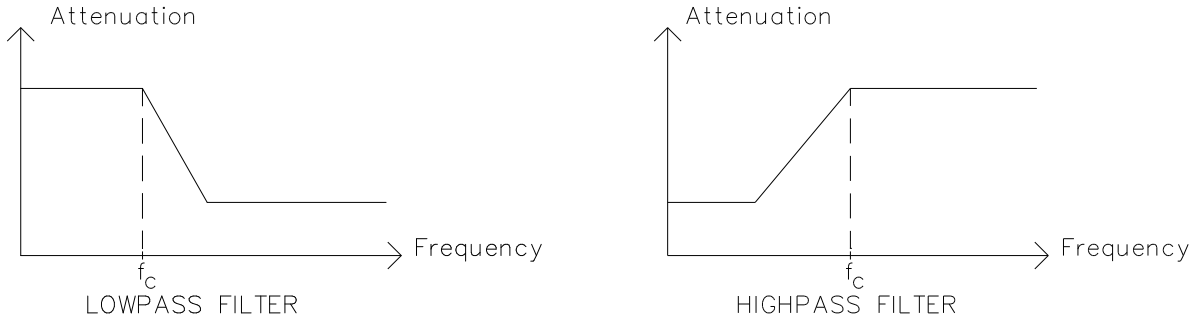


Figure 9-4: Two Commonly Used Analog Filters

To illustrate the need for bandlimiting speech, suppose a voice recording has been received with substantial hiss noise above 4 kHz. The signal-to-noise ratio (SNR) above 4 kHz is small, *i.e.*, very bad. The voice signal could be enhanced, *i.e.*, the overall voice SNR improved, by lowpass filtering the audio with a cutoff frequency F_c of 4 kHz. See Figure 9-5.

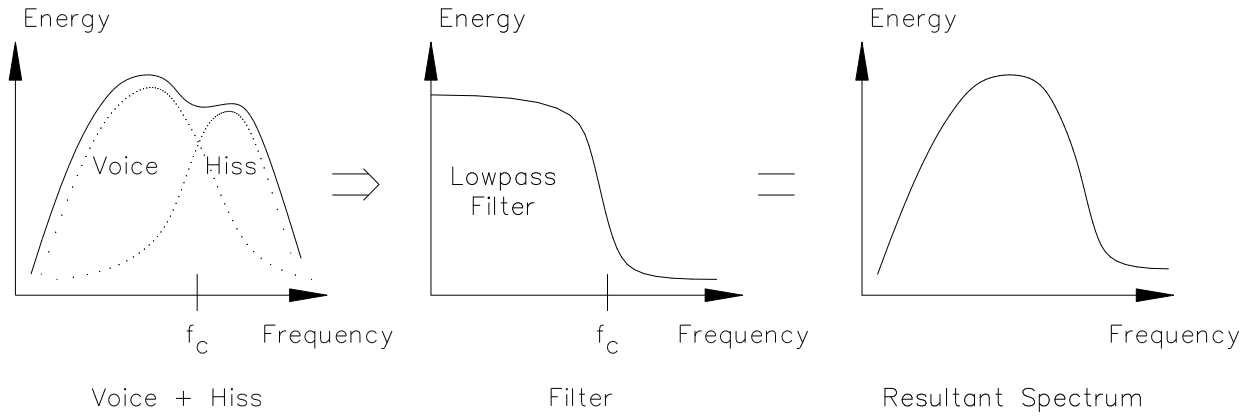


Figure 9-5: Lowpass Filtering

Though somewhat contrived, this example illustrates a method of voice enhancement by spectral amputation. Note that the high frequency components of the voice spectrum are attenuated in the process, but the cutoff frequency is high enough to retain the bulk of the voice information. Were F_c set lower to remove more hiss, additional, and perhaps unacceptable, voice energy would be lost.

The same approach can be used to remove low frequency room rumble from a voice using a highpass filter. This exercise is left for the reader.

Lowpass and highpass (as well as bandpass and bandstop) filters are characterized by the following parameters (or attributes):

- Cutoff frequency,
- Slope,
- Passband ripple,
- Stopband attenuation, and
- Phase characteristics.

The first four of these characteristics are illustrated in Figure 9-6.

The *passband* of the LPF is the frequency range over which energy is easily passed (in the figure, 0 Hz to 1000 Hz). The *stopband* is the range where energy is fully attenuated (in the example, 4000 Hz and up). The *transition band* (1000-4000 Hz) is the frequency interval where the filter is rolling off. The passband is said to end when the filter starts to roll off from its nominal value. The stopband starts when the specified attenuation is achieved.

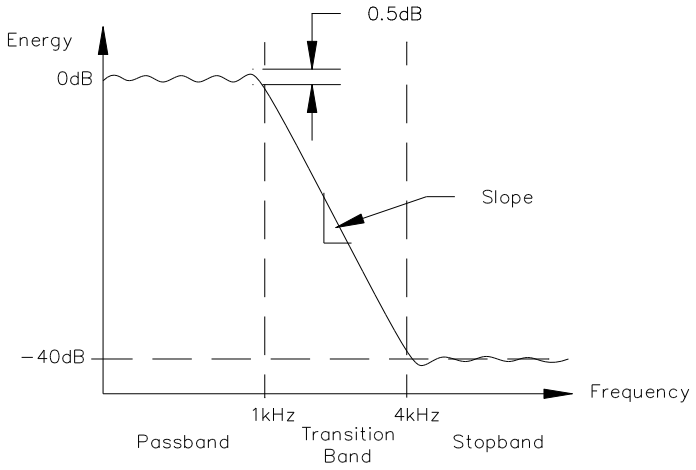


Figure 9-6: Lowpass Filter Characteristics Illustration

Under certain situations it is desirable to make the transition band as small as possible. This is called a "sharp cutoff" or "high Q" filter. Too sharp of a cutoff may cause the filter to *ring* (a slight oscillation or tinny sound occurring on impulsive and abrupt sounds).

The rate at which a filter rolls off in its transition band is its *slope* and in analog filters is expressed in dB per octave. In Figure 9-6, the filter rolls off 40 dB. Its slope is 20 dB/octave since two octaves are spanned. (As explained in Section 2.4, an octave is a doubling in frequency. 1 kHz to 2 kHz is one octave and 2 kHz to 4 kHz is the second octave, etc.) Digital FIR filters exhibit a uniform rolloff rate which is expressed as dB/Hz. For example, the characteristics shown in Figure 9-6 have a slope of 0.013 dB/Hz.

The *passband* ripple is the peak-to-valley variation of the passband characteristics. In the example, this is 0.5 dB. Ripple up to 1 dB is normally not objectionable for audio applications. Digital filters normally have very small ripple, and this parameter is often ignored.

The phase characteristics, the fifth parameter above, are somewhat esoteric. Phase describes how different frequency sounds are delayed through the filter. The ear is only mildly phase sensitive and phase characteristics are often not a priority in filter selection. The best audio phase characteristic is *linear phase*. A linear phase filter delays all frequencies exactly the same. Though linear phase analog filters (namely, Bessel-type) are possible, they have poor rolloff characteristics and are usually avoided in audio applications. Linear phase digital filters are easily implemented, however.

Figure 9-7 illustrates the digital lowpass filter control window of the PCAP[2]. The highpass control window is similar. These filters are fully *parametric*, meaning that the key parameters (cutoff

[2] The Personal Computer Audio Processor (PCAP) is a comprehensive audio filtering workstation manufactured by DAC. It has all the digital signal processing capabilities described herein, and thus the controls of the PCAP and PCAP II are used as examples.

frequency, slope, and stopband attenuation) are independently adjustable. Note that the slope is expressed as dB/octave even though the filter's slope is actually uniform (dB/Hz). The dB/Hz is converted to dB/octave, as this latter measurement is more familiar to most users.

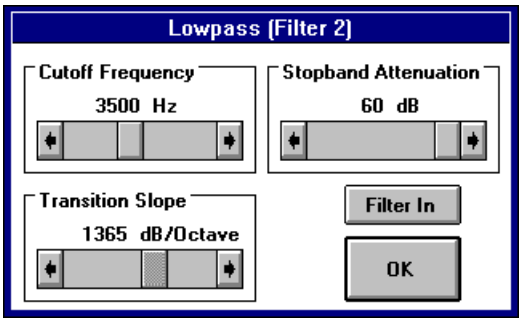


Figure 9-7: PCAP Lowpass Filter Control Window

Suppose both low and high frequency noises need to be removed simultaneously. This would require passing the audio through a lowpass filter to attenuate the high frequency noise *and then* passing the lowpass filtered audio through a highpass filter to attenuate low frequency noise. This is called *bandpass* filtering.

A bandpass filter (BPF) is the cascade (seriesing) of a lowpass and highpass filter. See Figure 9-8. Two cutoff frequencies are needed: F_L for the low cutoff frequency (corresponding to the LPF's F_C) and F_H for the high cutoff frequency. Though the terms "high," "low," "upper," and "lower" may be confusing at first, the reader is encouraged to develop a comfortable intuitive understanding of these concepts.

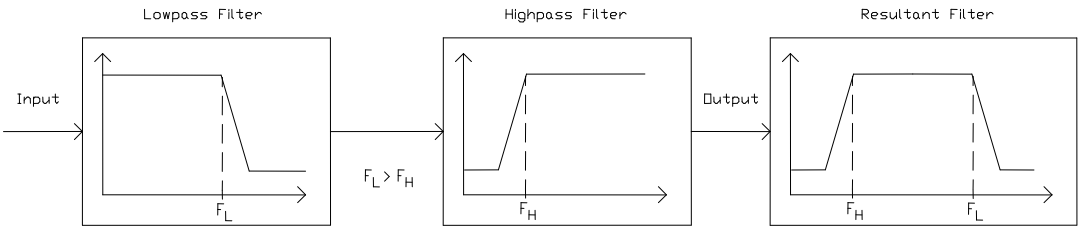


Figure 9-8: Bandpass Filter (BPF)

The converse to a bandpass filter is the *bandstop filter* (BSF), also known as a band-reject filter. As illustrated in Figure 9-9, a bandstop filter is produced by highpass filtering and lowpass filtering the *same* input audio and the summing (adding) the two filters' outputs. A BSF passes energy below F_L (from the LPF) and above F_H (from the HPF) and attenuates energy between F_L and F_H. BSFs are used to remove midband noise.

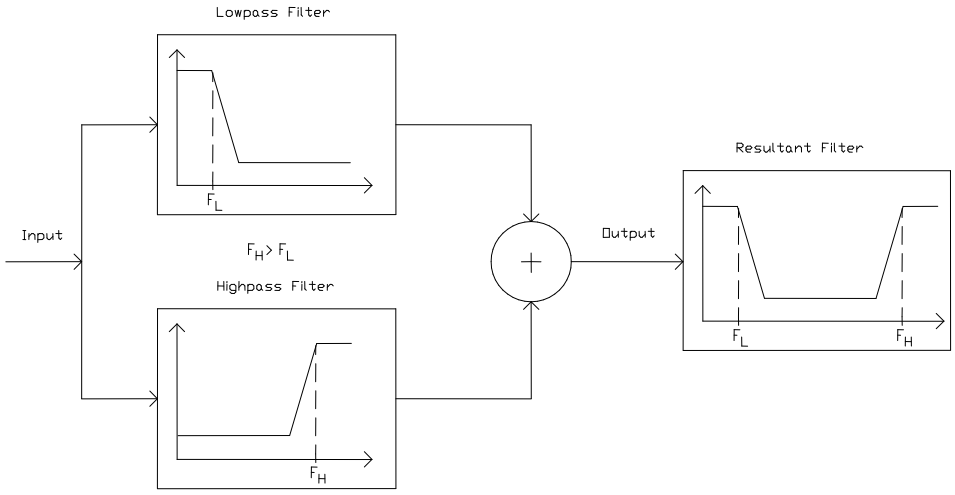


Figure 9-9: Bandstop Filter (BSF)

Figure 9-10 illustrates the PCAP's digital bandpass filter control windows. Like other filters in the PCAP, this one also allows independent adjustment of filter parameters (See Figure 9-11). The bandstop filter control window and parameters are similar.

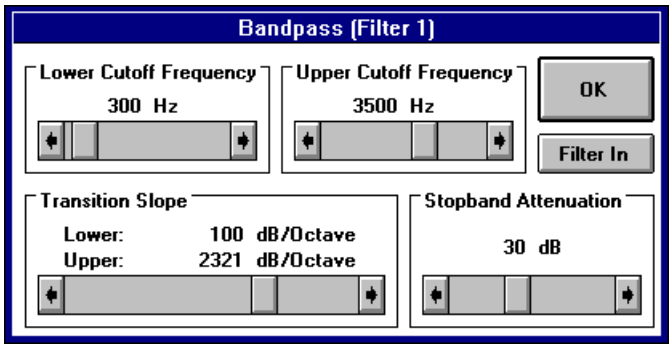


Figure 9-10: PCAP Bandpass Filter Control Window

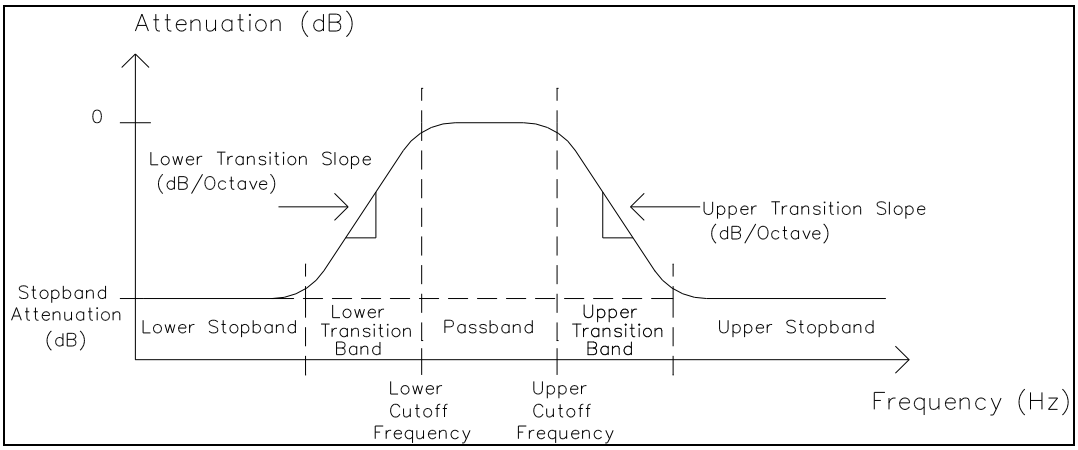


Figure 9-11: Bandpass Filter Parameters

9.2.2 Notch and Slot Filters

A very narrow bandstop filter is called a *notch filter*. Notch filters are used to remove tones and narrow-band noises from audio with minimal amputation of the desired signal. Unlike the bandpass filter, the lower and upper cutoff frequencies (F_L and F_H) are normally not specified, but instead the *center frequency* (F_C) and *bandwidth* (BW) are. An additional distinction between these filters is that a notch filter's stopband tapers to a point of maximum attenuation, whereas a bandstop filter's stopband is usually flat. Consider the notch filter in the following illustration.

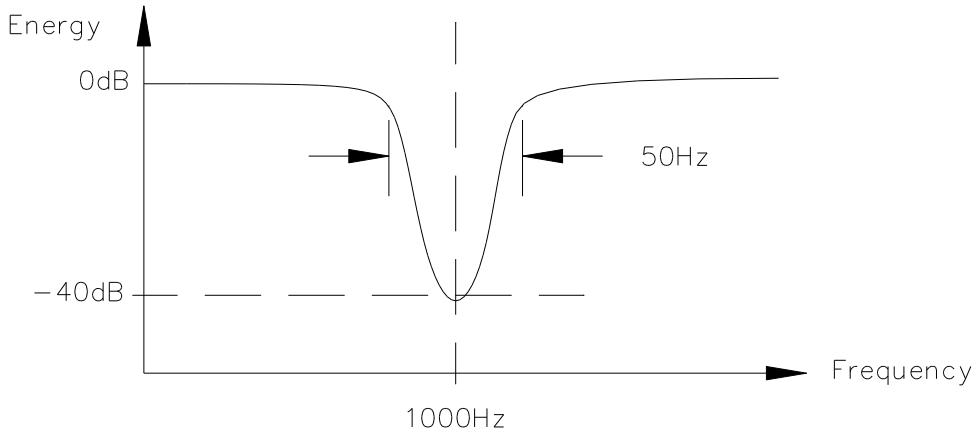


Figure 9-12: Notch Filter Illustration

The notch filter has a center frequency of 1 kHz and a very narrow bandwidth of 50 Hz. If F_L and F_H were used to describe this filter, F_L would be 975 Hz and F_H would be 1025 Hz. Its attenuation is 40 dB.

A *slot* filter is the opposite of a notch filter. It is a narrow bandpass filter which passes the audio signal in only a narrow band. Like the notch filter, center frequency F_C and bandwidth BW are specified. Slot filters are useful in isolating narrow signals such as background tones and dialing tones.

In analog filters, the *quality factor*, or Q, of a notch or slot is usually specified. Q is defined as

$$Q = F_C / BW$$

and is large for narrow filters. (Q is also sometime used to describe the sharpness of a lowpass or highpass transition band.)

> Example: What is the Q of a notch filter whose stop band goes from 2050 Hz to 2300 Hz?

$$BW = 2300\ Hz - 2050\ Hz = 250\ Hz$$
$$F_C = ½\ (2300\ Hz + 2050\ Hz) = 2175\ Hz$$
$$Q = 2175\ Hz / 250\ Hz = 8.7$$

Figure 9-13 illustrates the PCAP's digital notch filter control window. This is also a parametric filter, and independent adjustment of notch center frequency, notch width, and notch depth are provided. The width is displayed rather than Q, common in analog filters, as it is more meaningful. Digital notch filters are capable of achieving much narrower notches than their analog counterparts.

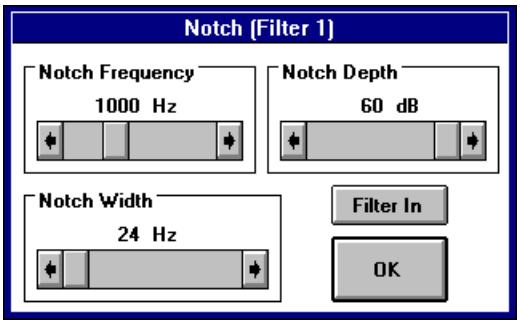


Figure 9-13: PCAP Notch Filter Control Window

Analog notch and slot filters are universally nonlinear phase. Digital notch and slot filters may be either linear or nonlinear phase.

9.2.3 Spectrum Equalizers

Audio processing makes use of spectrum equalizers to readjust overall spectral shape and, to a limited extent, reduce banded noises and tones. Three commonly-used types of equalizers are the *multiband graphic equalizer*, *spectral graphic equalizer*, and *parametric equalizer.*

9.2.3.1 Multiband Graphic Equalizer

Of the three types of spectrum equalizers, the easiest to operate and understand is the graphic equalizer, functionally illustrated in Figure 9-14.

A multiband graphic equalizer partitions the audio spectrum into 15 to 25 segments using bandpass filters. It adjusts the relative gain of each segment with panel slide-controls and then combines the resulting gain-adjusted spectral segments into a single output. The spectral transfer function, *i.e.*, how the output signal spectrum is changed from the input signal spectrum, is graphically displayed by the position of the slide controls on the instrument's front panel.

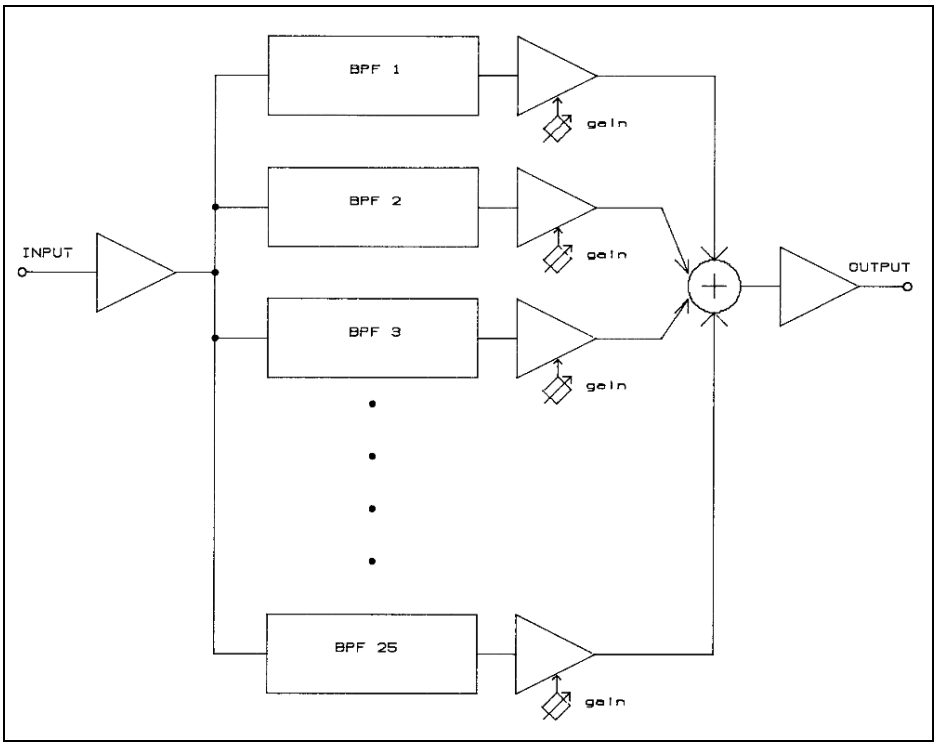


Figure 9-14: 20-Band Graphic Equalizer

It is very easy to use the graphic equalizer and to observe its spectral shape (by looking at the slide control positions). The principal application of an *analog* graphic equalizer is in sound

reinforcement (PA) systems and music spectral shaping. Unfortunately, it has three serious drawbacks when applied to forensic voice noise filtering: poor bandpass filter resolution, sensitivity to impulsive noises, and inability to produce sharp notches.

Analog graphic equalizers usually incorporate one-third octave or octave bandwidths BPFs. These bandpass filters become progressively wider as the center frequency becomes higher. As a result, the low frequency resolution is a few tens of hertz wide, whereas the high frequency resolution is several hundreds of hertz wide. This arrangement is acceptable for sound reinforcement systems but not for noise filtering of voice signals.

Constant bandwidths, *i.e.*, all bandpass filters having the same bandwidths, are not available in analog graphic equalizers due to the limitations of analog circuit components. Digital filters, however, are able to implement constant bandwidth configuration, typically 175 to 350 Hz.

The second drawback of analog graphic equalizers is that they behave poorly in the presence of impulsive noise. Note that the output signal is the sum of 15 to 25 bandpass filters. Each filter has sharp skirts, *i.e.*, rapid cut off. Such filters have a high Q factor and tend to ring when driven by an impulse. The output is then the sum of 15 to 25 bandpass filter ringings, an undesired artifact produced by the instrument. Digital equalizers, however, use a single linear phase filter, which virtually eliminates ringing.

Digital graphic equalizers, like analog equalizers, are unable to produce sharp notches at arbitrary frequencies. By combining a notch filter (or multinotch filter) with a digital graphic equalizer, all requirements are met. The principal advantage of this approach is ease of adjustment. The overall equalization characteristics of a constant-bandwidth digital equalizer are given as the dashed curve in Figure 9-15. This curve is the composite effect of the individual bandpass filters and their respective gains.

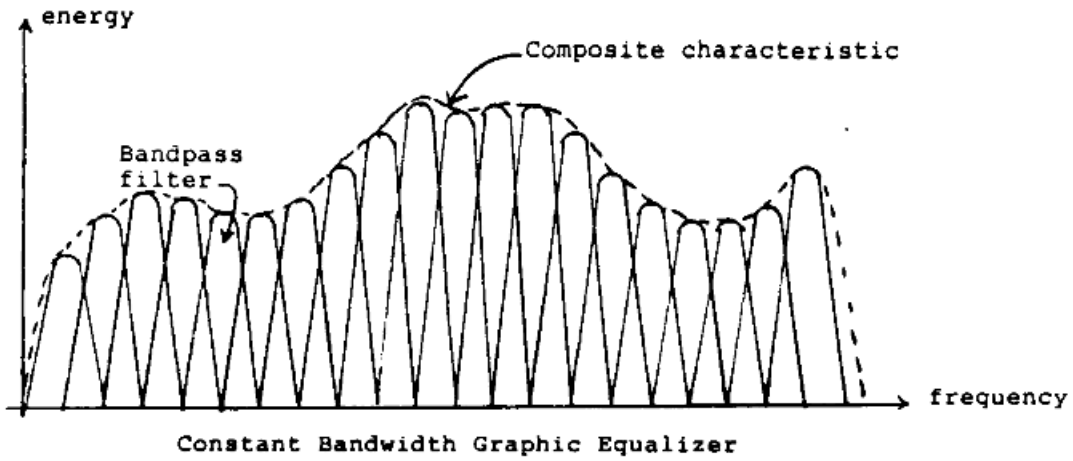


Figure 9-15: Graphic Equalizer Curve

Figure 9-16 illustrates the PCAP's 20-band digital graphic equalizer control window. This software-controlled equalizer offers storage and recall of equalization curves from memories and allows moving all slide controls down or up collectively.

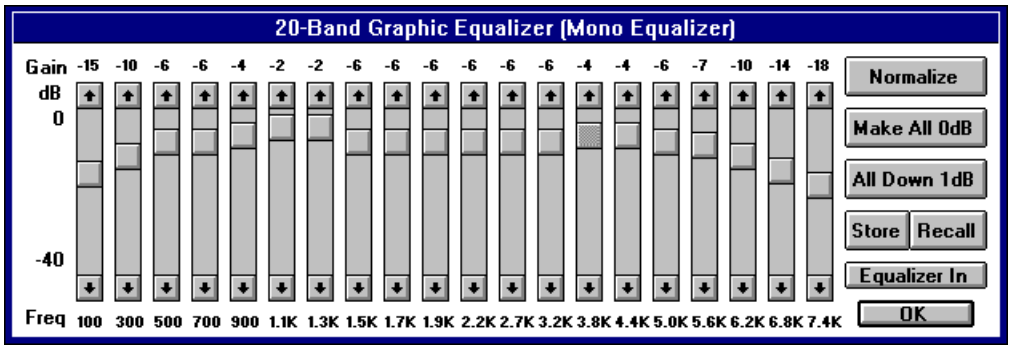


Figure 9-16: 20-Band Graphic Equalizer

9.2.3.2 Spectral Graphic Equalizer

The spectral graphic equalizer is a very-narrow-band graphic equalizer. As shown in Figure 9-17, the PCAP's Spectral Graphic Equalizer has 115 bands. Instead of adjusting each band separately, an equalization curve is mouse drawn. Each band's slide control is placed at its appropriate position on the drawn curve automatically by the software.

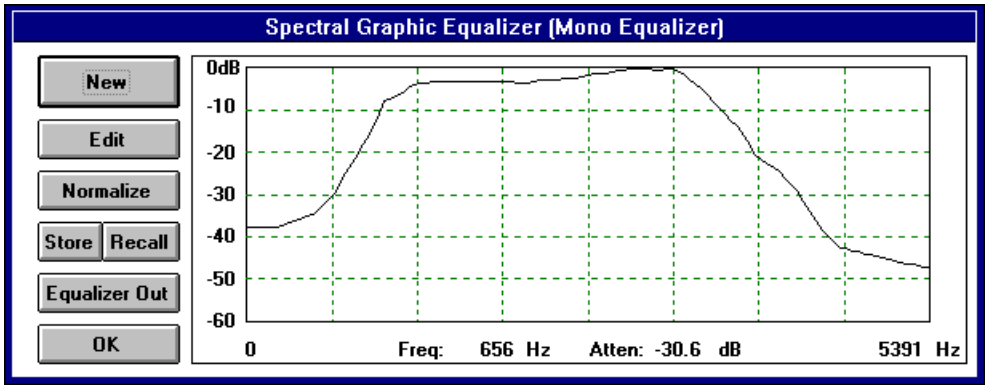


Figure 9-17: PCAP Spectral Graphic Equalizer Control Window

Each band has uniform width. When the PCAP is configured for 5.4 kHz processing bandwidth, each band is

$$5400 \text{ Hz} / 115 = 47 \text{ Hz wide.}$$

This resolution permits the spectral graphic to simultaneously achieve sharp notches and spectral shaping. The spectral graphic equalizer has memories for storing and recalling previously-drawn equalization curves and the ability to edit an existing curve.

9.2.3.3 Parametric Equalizer

The parametric equalizer is functionally illustrated in Figure 9-18.

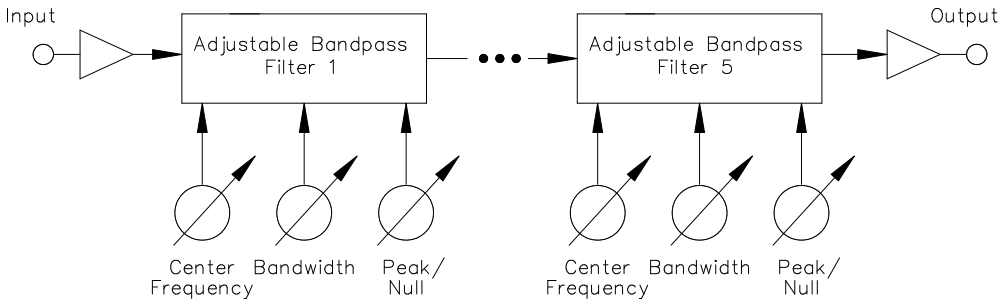


Figure 9-18: Functional Block Diagram of a Parametric Equalizer

The analog parametric equalizer has been widely used for forensic voice processing. This instrument offers substantial control over the spectral shape and allows relatively good control for removing bandlimited noise and tones. The parametric equalizer requires a reasonable level of skill to operate, and it normally is used in conjunction with an FFT spectrum analyzer.

Parametric equalizers consist of four to eight tunable bandpass filters in cascade. Each filter has three adjustable parameters. These are *center frequency*, *bandwidth* or Q, and *boost/cut* gain. The center frequency and bandwidth control the overall shape of the individual equalizing filters. Typically, the center frequencies may be adjusted over the range of 80 Hz to 6000 Hz and the bandwidth from 3.4× to 0.2× the center frequency. At 0.2× bandwidth and a 1000 Hz center frequency, a 200 Hz wide filter is produced. (The bandwidth is normally expressed as the Q factor. The above range corresponds to a Q range of 0.29 to 5.)

$$Q = F_C / BW$$

The third control (parameter) is the boost/cut gain control. A differential amplifier is used to permit the bandpass filter to either intensify (boost) or attenuate (cut) the overall spectrum. Attenuating will produce a notch in the transfer function. Intensifying will elevate low energy sections of the audio spectrum.

Figure 9-19 shows the effect of adjusting both the Q and the boost/cut gains of a single bandpass filter section. The center frequency is set to 1000 Hz. Similar curve sets may be produced at other center frequencies.

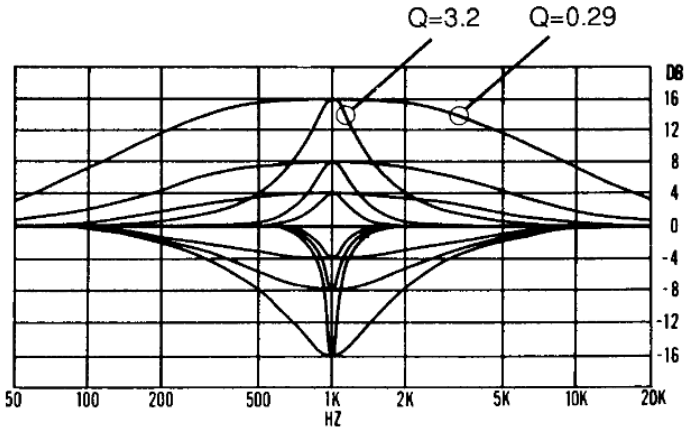


Figure 9-19: Typical Boost/Cut Curves
(with center frequency and bandwidth fixed)

Figure 9-20 shows the PCAP II's parametric equalizer control window. It uses the three previously-described parameters to control each filter substage. Note that the width is controlled by specifying bandwidth in hertz rather than the Q factor.

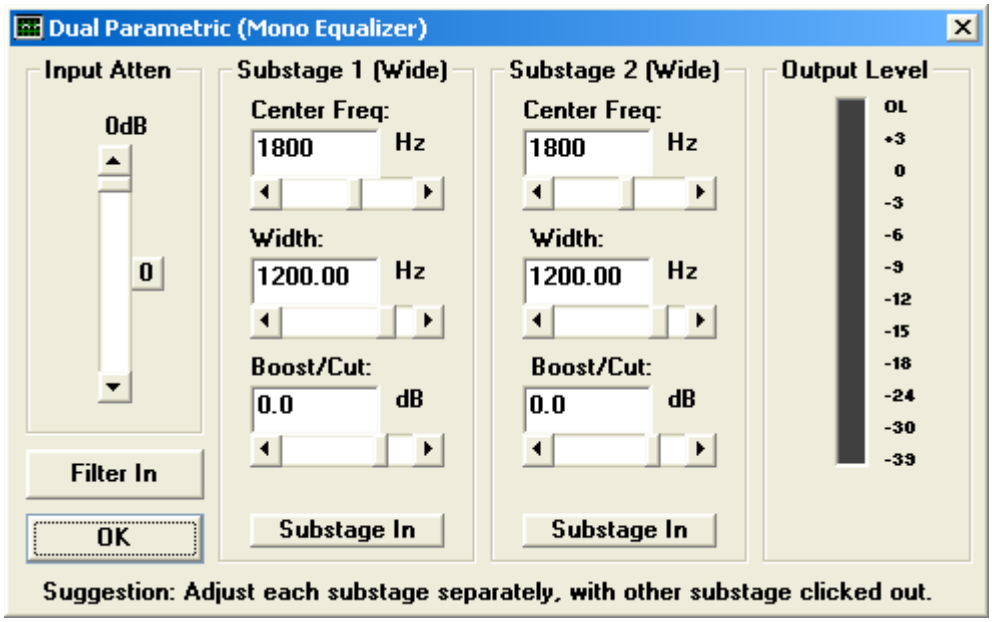


Figure 9-20: PCAP II Parametric Equalizer Control Window

Figure 9-21 illustrates the overall characteristics (called *transfer curve*) of a four-section parametric equalizer. Individual bandpass filter sections 1 through 4 are combined to produce the overall equalization transfer curve.

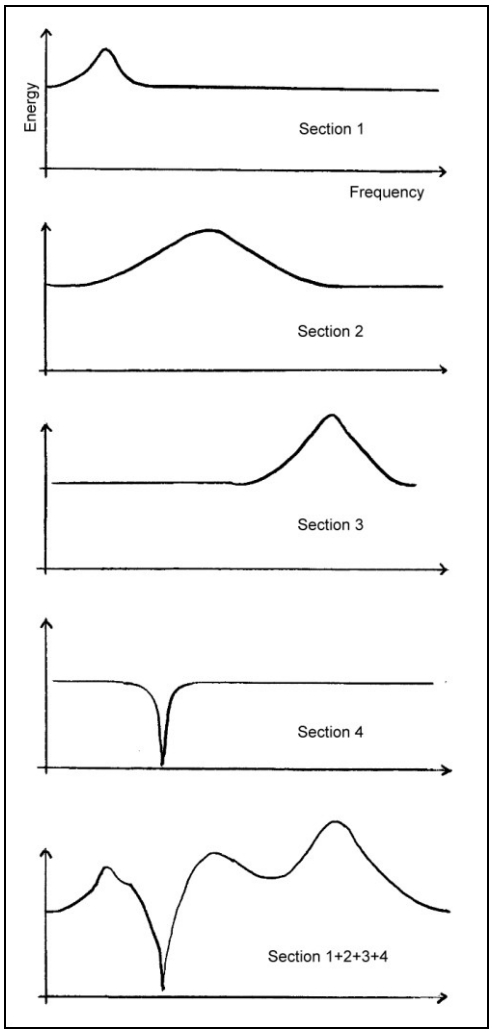


Figure 9-21: Parametric Equalization Curve

By having adjustable equalizing filters rather than fixed proportional bandwidth (third octave) bandpass filters, the parametric equalizer gives substantially greater control over spectral shape compared to an analog graphic equalizer. Since it has many fewer filters and the skirts of these filters can be controlled, ringing effects in the presence of impulsive noise is also reduced. The key disadvantage of the parametric equation is that adjustment requires an appreciable level of skill, and an FFT spectrum analyzer is normally required.

9.2.4 Analog Filter Implementations

This section discusses analog filter characteristics. Digital filters may also be designed with these same characteristics; additional, and more robust, filter types also are possible using digital filters. Analog filters are designed to achieve different goals (called design criteria). The vast majority of analog filters fall into one of four filter types given in Table 9.

Table 9: Filter Design Criteria

Butterworth	· Maximally flat amplitude · Smooth phase response · 6 dB/octave/pole rolloff
Bessel Filter	· Constant delay at all frequencies (linear phase) · Flat amplitude response · 3 dB/octave/pole rolloff
Chebyshev	· Specified ripple in passband · Nonlinear phase · Rapid, monotonic rolloff
Elliptic	· Specified ripple in passband · Specified stopband depth · Nonlinear phase · Rapid rolloff

9.3 Dynamic Audio Level Control

Three devices are commonly used for automatically controlling audio levels on forensic recordings:

- Limiter,
- Automatic Gain Control (AGC), and
- Compressor/expander.

All three use electronically-controlled amplifiers and various types of signal-level sensing circuits. An electronically-controlled amplifier (ECA) is an amplifier whose gain (amplification) is controlled by the audio level itself. Instruments produced by Alesis, dBX, and DAC are commonly employed for these functions.

9.3.1 Limiter

An audio limiter is a device which attempts to keep audio levels from exceeding a specific threshold level. Such devices are very useful in preventing accidental overload of input circuits from abrupt sounds such as door slams. Should a signal peak approach this threshold, the electronically-controlled amplifier (ECA) gain is automatically reduced. If no further loud audio is experienced, the gain slowly returns to its pre-limiting value. Note that a limiter is not a clipper, which merely chops off the audio peaks introducing nonlinear distortion.

Figure 9-22 functionally illustrates a limiter circuit. This circuit consists of an electronically-controlled amplifier, an audio peak level detector, and a gain control circuit. Limiters normally have input and output amplifiers to match the audio levels of the input and output instruments; for simplicity, these are not shown in the figure.

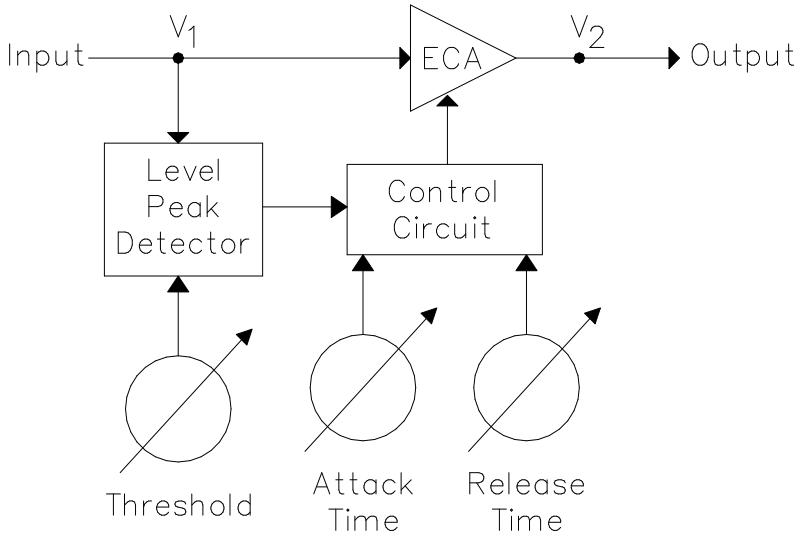


Figure 9-22: Limiter Circuit

The ECA does not amplify but acts as a controlled *attenuator*. Under normal conditions, there is no attenuation (gain = 1) and $V_1 = V_2$. If loud audio appears at the input and exceeds the threshold, the control circuits instruct the ECA to attenuate (gain < 1) V_1 such that V_2 does not exceed the threshold. This automatic reduction in gain occurs very rapidly, on the order of a few milliseconds. This *attack time* may or may not be adjustable, depending upon the instrument.

Once the loud audio disappears, the ECA gain is allowed to slowly return to normal (gain = 1). The period required for the return is adjustable as the *release time*. Figure 9-23 illustrates the gain curve of a limiter. In this case the threshold is set to 0 dB. Audio levels below 0 dB have a constant gain. The 45° gain segment has the output increasing 1 dB with each 1 dB increase in input levels. If a loud input signal exceeds 0 dB, the limiter's output remains at 0 dB.

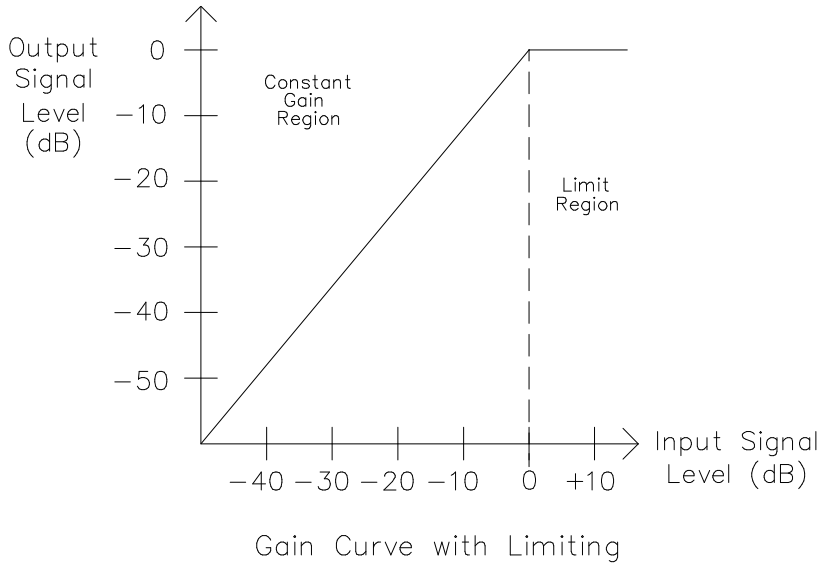


Figure 9-23: Limiter Gain Curve

Limiters are used where there is a possibility of input signal overload due to unpredictable loud sounds. Tape recorders often have such circuits built in.

Figure 9-24 shows the limiter control window from the PCAP. The attack time is very fast (approximately 1 msec) and is not adjustable. The release time is selected as 250 msec. This limiter will rapidly drive down its gain automatically in the presence of a loud signal, *i.e.*, on having a level greater than –9 dB (the threshold) on the PCAP's input bar graph. Once the audio level falls below –9 dB, the gain will return to normal in approximately 250 msec.

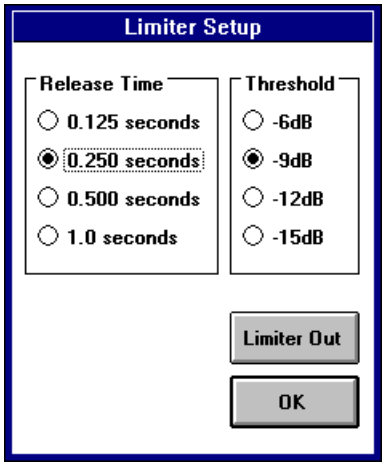


Figure 9-24: Limiter Control Window

9.3.2 Automatic Gain Control

An automatic gain control (AGC), also called an automatic level control, is a device which attempts to maintain all audio peaks at a uniform level. Both soft and loud voices are maintained at a uniform level with an AGC. Inexpensive cassette tape recorders use AGC circuits to avoid the need for a record level control. Forensic telephone recordings with near/far party volume levels also benefit from an AGC.

The AGC is functionally similar to the limiter; the main difference is the control logic. An AGC is an amplifying (gain > 1) device; whereas, a limiter is an attenuating (gain < 1) device. Figure 9-25 functionally illustrates an AGC.

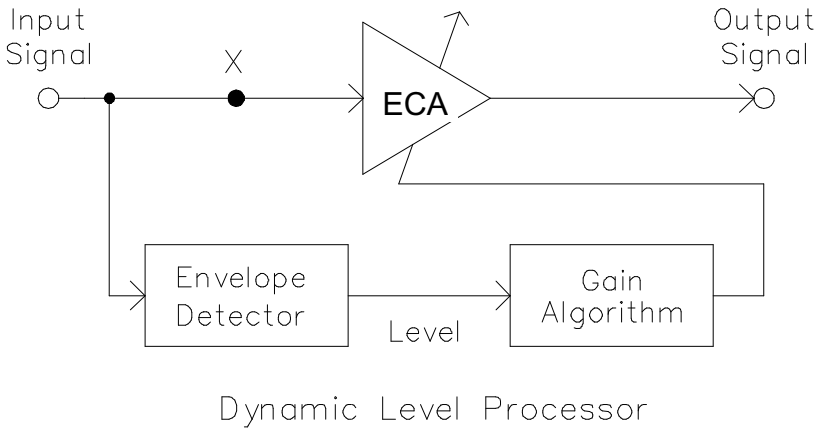


Figure 9-25: AGC Functional Block Diagram

The AGC usually incorporates two controls: maximum allowable gain and release time. The AGC maximum gain specifies the amount of amplification, or *boost*, that may be applied to a weaker talker in an attempt to elevate its volume to that of the louder audio. Boost levels of 10 to 20 dB are commonly used. During silences, background noise is also boosted, since AGCs cannot distinguish between voices and background noises. Too great a boost causes annoying noisy intervals between speech parcels.

Unlike a limiter, the AGC's ECA normally operates at a gain greater than 1; the majority of the audio is not loud peaks. When a peak occurs, the ECA must rapidly reduce its gain in order to avoid excessive output levels. The time interval required for an AGC to respond to loud sounds is its *attack time*, which is usually not adjustable and is typically a few milliseconds. Once the loud sound has disappeared, the ECA attempts to return to a high gain in order to elevate low level sounds. The time required for the ECA's gain to build back up is the *release time* and is typically 100 to 1000 msec.

In order to avoid overload on loud peaks, the AGC's attack time must be made very short. Short attack and release times cause the ECA's gain to change rapidly causing an annoying *pumping* sound. A digital AGC can avoid this problem by placing a short digital audio delay line ahead of

the ECA, at "X" in Figure 9-16; the level detector and control circuitry can then see a loud signal before it reaches the ECA. The attack time can be increased since the control circuitry can now *anticipate* the need for gain reduction. This look-ahead feature thus allows gain control without the pumping effects of fast AGCs.

Figure 9-26 is the PCAP's AGC control window. This device has a rapid attack time of approximately 1 msec and a selectable release time, as shown, of 200 msec. The PCAP's AGC maximum boost in gain is 20 dB.

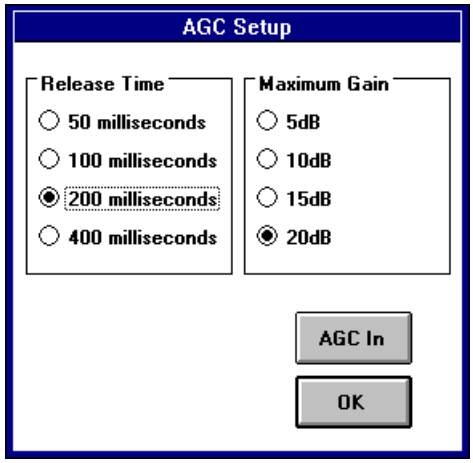


Figure 9-26: PCAP AGC Control Window

Using this device with a loud talker and a soft talker at a 10 dB lower voice level, the AGC will instantaneously (1 msec) reduce its gain 10 dB when the conversation switches from the lower to the louder voice. When the conversation switches back to the lower voice, the gain is increased 10 dB but more gradually over a period of 200 msec. When neither talker is speaking, the background noise is amplified the maximum gain of 20 dB.

9.3.3 Compressor/Expander

A compressor/expander is very useful in normalizing the levels of multiple voices and suppressing background noise in nonvoice segments of a recording. Such a device normally has two regions of operation. The *compression region* squeezes voices into more limited output level range, thereby making all voices sound similar in loudness. The *expansion region* reduces the level of sounds below the voice levels.

The three main controls of a compressor/expander are

- compression ratio,
- expansion ratio, and
- compression threshold.

The ratios relate the change in output level to the change in input level. Consider two voices, one at a –10 dB level and one at +5 dB level. Their difference is

$$\text{input level difference} = +5 - (-10) = 15 \text{ dB}$$

If a 3:1 (or 3.0) compressor is applied to this audio, then the resultant difference becomes 5 dB, *i.e.*,

$$\text{compression ratio} = \frac{\text{input level difference}}{\text{output level difference}}$$

The expansion ratio is actually a compression ratio with a value less than 1, *e.g.*, 1:2. If the two voices above were expanded by an expansion ratio of 1:2 (or 0.5), then differences would be 30 dB. Note that an expansion ratio (or compression ratio) of 1:1 (or merely 1.0) is no expansion or compression at all. The output level differences are the same as the input level differences.

The compression threshold is the audio level where expansion stops and compression starts. In the voice example above, this threshold would be set *below* the softest voice, *e.g.*, –15 dB. The voice would be compressed at a ratio of 3:1 and during silence the background noise would be suppressed by stretching it lower at an expansion ratio of 2:1. Figure 9-27 illustrates.

Compressor/expanders are combined with a limiter and have a limiter threshold. This control limits the maximum output level by placing a cap on output audio independent of the input level. In the example, the limiter threshold would be set at a value greater than the loudest voice, *e.g.*, +10 dB. A limiter is a compressor with a ∞:1 compression ratio.

A compressor is similar to an AGC except that not all signals are amplified to the same level. Instead, the amplitude differences between loud and soft sounds are reduced. Consider two voices where one has a level of 0 dB and the other –10 dB. If a 2:1 compressor is used, the louder sound would remain at 0 dB but the softer would be compressed to –5 dB. A 4:1 compression ratio would reduce the amplitude difference to 2.5 dB, and a 100:1 compression would virtually eliminate any differences in the two voice levels. A large compression ratio makes a compressor become an AGC.

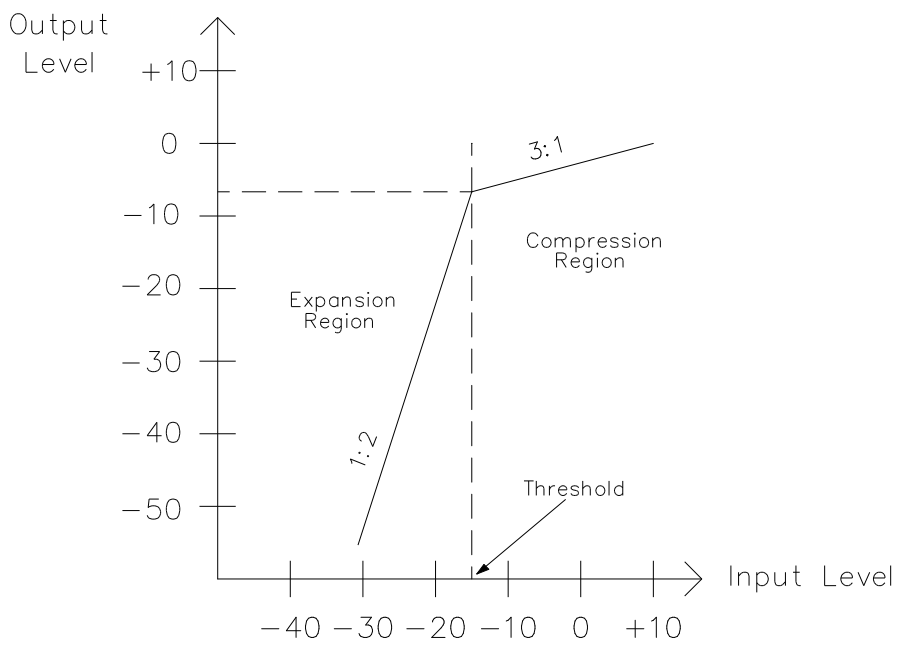


Figure 9-27: Compression/Expansion

The main advantage of a compressor over an AGC for forensic applications is that the difference in voice levels is reduced in a more controlled fashion. Voice levels are still distinguishable, but softer and louder voices are brought closer in volume. Compression ratios greater than 2:1 may start to sound unnatural.

Like AGCs, compressors also have attack and release times used by signal level measurement circuits to measure the signal level.

EXERCISES

1. What type of bandlimited filter is commonly used to reduce low-frequency room noises? High frequency tape hiss?

2. A 1 kHz lowpass filter has a rolloff rate of 12 dB/octave.
   How much attenuation would exist at 500 Hz, 2 kHz, 8 kHz, and 16 kHz?

3. A 1 kHz highpass filter has a rolloff rate of 12 dB/octave.
   How much attenuation would exist at 250 Hz and 62.5 Hz?

4. A lowpass filter has a cutoff frequency of 1 kHz and rolls off monotonically (continuously). At 4 kHz the attenuation is –48 dB. What is its rolloff rate in dB/octave?

5. Which type of analog filter (Section 9.2.4) has the least phase distortion?
   The second best phase characteristics?

6. A four-pole Butterworth highpass filter has a cutoff frequency of 1 kHz.
   What are its attenuations at 250 Hz, 500 Hz, and 2 kHz?

7. What are the three most important parameters of an audio bandlimit filter?

8. A notch filter has a center frequency of 1 kHz and a Q of 10.
   What is its notch width? What are its upper and lower cutoff frequencies?

9. A compressor/expander has two voices centered at –10dB and –25dB, respectively. The compression threshold is –40dB, and the compression ratio is 3:1. What is the level difference before and after compression?

10. DIGITAL SIGNAL PROCESSING

The advent of high-speed microprocessor technology has benefited a multitude of technological areas including signal processing. Digital signal processing (DSP) consists of three principal components, illustrated in Figure 9-3 of Section 9.1.

Analog-to-digital (A/D) conversion is the process whereby input analog audio is converted to a stream of binary numbers, each of which represents a voltage measurement, or *sample*, of the audio waveform at a specific point in time. Digital-to-analog (D/A) conversion is the inverse process whereby binary samples are reconverted to analog audio which is suitable for amplification and hence listening. The digital computer component is a very-high-speed numeric processor for implementing the DSP mathematical equations.

In order for such powerful instruments to be effective in forensic voice processing, they must both produce the desired results and be easy to operate.

10.1 Audio Sampling

An analog-to-digital converter, or ADC, encodes an analog signal into a series of digital representations. A digital-to-analog converter, or DAC, decodes a digital signal to form a corresponding analog signal.

10.1.1 Analog-to-Digital Conversion

Audio sampling is the process by which instantaneous voltage measurements are made on the audio waveform. These measurements are usually made by a sampling analog-to-digital converter. This sampling process must conform to the *Nyquist sampling theorem* in order to properly encode the analog waveform.

The Nyquist sampling theorem states that

> For an analog signal to be properly sampled, instantaneous voltage measurements must be made on that signal at a rate at least twice F_N, where no significant energy in the signal exists above the frequency F_N.

The frequency F_N is commonly called the *Nyquist frequency*, and $2F_N$ is called the *Nyquist rate*.

Suppose an audio signal is measured with an FFT spectrum analyzer and found to have all of its energy contained below 5 kHz. The Nyquist theorem says that the minimum sampling frequency for that signal is $2 \times F_N$ or 10 kHz. Actually, any sampling frequency above 10 kHz may be

used; however, excessively high sampling frequencies yield no benefit but do add additional processing burdens. For this reason, excessive oversampling is usually avoided.

In order to assure that no signal energy exists above F_N, an analog lowpass filter is normally used. Lowpass filters cannot roll off instantaneously, which means that they must transition over a range of frequencies rather than at a single frequency; thus, the actual sampling frequency is usually greater than twice the audio bandwidth. Since lowpass filters must transition over a range of frequencies rather than at a single frequency, cannot roll off instantaneously, the actual sampling frequency is usually greater than twice the audio bandwidth. Consider the sharp lowpass filter in Figure 10-1 that passes audio up to 5 kHz but does not roll off completely until 6.25 kHz. Energy exists in the 5 kHz to 6.25 kHz transition band. Nyquist specifies that the *minimum* sampling frequency for this configuration is 12.5 kHz. If a signal is undersampled with a sampling frequency less than the Nyquist rate, a phenomenon called *aliasing distortion* occurs. This nonlinear distortion results in the shifting and blending of spectral components, usually at high frequencies. Using an adequate sampling frequency avoids this problem.

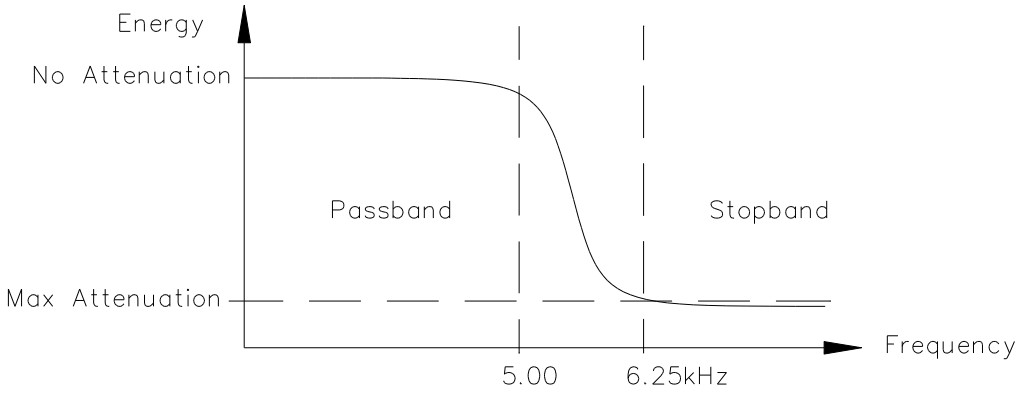


Figure 10-1: Sharp Cutoff Lowpass Sampling Filter

The input lowpass filter in the sampling system of Figure 9-3 is often called an *antialiasing* filter. This filter limits the audio bandwidth and thus permits specifying the minimum sample rate. This filter prevents aliasing distortion, which occurs when the signal's bandwidth exceeds half the sampling frequency. Such a filter is usually flat in its passband and has a specified attenuation in the stopband. If the sampling system has a required dynamic range of 60 dB, then the stopband should have an attenuation of at least 60 dB. This assures that any aliasing distortion components will be below the dynamic range of the system.

The final stage in the analog-to-digital conversion process is the sampling analog-to-digital (A/D) converter. See Figure 10-2. This digital voltmeter makes instantaneous voltage measurements on the lowpass filtered, or *bandlimited*, audio and outputs these binary measurements to the digital computer. A/Ds perform a sequence of steps in order to measure the voltage. This process is fast but still takes time. In order to meet the instantaneous sampling requirement, a sample-and-hold amplifier (S/H) is usually placed ahead of the A/D. This amplifier takes an instantaneous snapshot

of the voltage and holds that value long enough for the A/D to make its measurement. Most new A/D designs are sampling analog-to-digital converters and have an S/H built in.

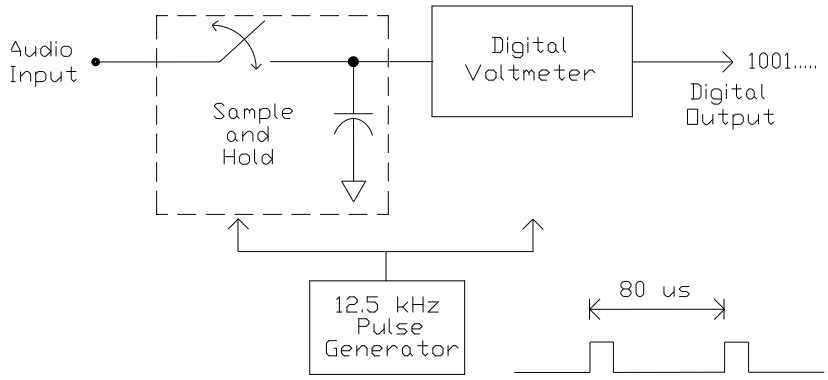


Figure 10-2: Sampling A/D

The A/D receives a synchronous strobe signal, or sample clock, which instructs it to make the measurement. The strobe, in the example above, occurs every

$$\frac{1}{12{,}500\ \text{Hz}} = 80 \text{ microseconds (µsec).}$$

The resolution of the A/D is measured in bits. Each bit doubles the dynamic range, which corresponds to a 6 dB[3] increase in resolution. The dynamic range is thus specified as

$$\text{dynamic range} = n \times 6 \text{ dB},$$

where n is the number of A/D bits of resolution.

As an example, a 16-bit A/D, commonly used in CD players and DAC filters, has a theoretical dynamic range of 16 bits × 6 dB/bit = 96 dB.

A/D converters utilize *binary* arithmetic. Table 10 below lists several 8-bit numbers and their decimal values. Note that there are a total of $2^8 = 256$ eight-bit numbers possible. If 10 bits are used to represent the number, then $2^{10} = 1024$ different 10-bit numbers are possible.

[3] $20 \log\left(\frac{2}{1}\right) = 6.02$ dB

Table 10: Examples of Eight-Bit Binary Numbers

Decimal Number	Equivalent Binary Number
0	0000 0000
1	0000 0001
↓	↓
5	0000 0101
↓	↓
16	0001 0000
↓	↓
128	1000 0000
↓	↓
254	1111 1110
255	1111 1111

A *Successive Approximation Register* (SAR) analog-to-digital converter converts a voltage using a sequence of steps, each step halving its range and narrowing down on the actual value. The following illustration shows how an SAR A/D gives a binary measurement for a 2.6 volt sample (*V*) of audio.

Example of analog-to-digital conversion:

The A/D has a range of 0 to 4 volts.

Step 1: Is *V* greater or less than 2 volts?
2 volts is midway between the current range of 0 to 4 volts.

Answer: Greater; therefore, the first bit is 1. The current binary number is 1???.

Step 2: Is *V* greater or less than 3 volts?
V is between 2 and 4 volts, and its midpoint is 3 volts.

Answer: Less; therefore, the second bit is 0. The current binary number is 10??.

Step 3: Is *V* greater or less than 2.5 volts?
V is between 2 and 3 volts, and its midpoint is 2.5 volts.

Answer: Greater; therefore, the third bit is 1. The current binary number is 101?.

Step 4: Is *V* greater or less than 2.75 volts?
V is between 2.5 and 3 volts, and its midpoint is 2.75 volts.

Answer: Less; therefore, the fourth bit is 0. The final binary number is 1010.

Obviously, this process can be carried out for additional steps. Each step improves the precision of the binary number and adds 6 dB of dynamic range.

Recent developments in analog-to-digital conversion have resulted in *sigma-delta oversampling* converters. This technique reduces the analog lowpass filter requirements by using more robust on-chip digital filtering. The analog lowpass filters do not need sharp cutoffs and are typically implemented as simple resistor-capacitor (RC) pairs.

The sigma-delta A/D greatly oversamples the input audio at rates of 64 to 256 times F_N using a 1-bit A/D. An on-chip microprocessor continually filters the results of these conversions and forms 16- to 24-bit values at a rate of 2 F_N. These values are very high quality, achieving near-ideal theoretical performance in terms of their numeric precision and accuracy. Sigma-delta 24-bit A/Ds are the current state of the art and are used on all current-generation DAC products.

10.1.2 Digital-to-Analog Conversion

The inverse process to analog-to-digital conversion is digital-to-analog conversion. This process reconstructs the analog signal from the digital samples. The stream of digital values produced by the digital computer is reconverted to analog samples by the digital-to-analog (D/A) converter. This device is a digitally-controlled voltage source. Every 80 µsec, in the example above, the D/A outputs a new voltage level. Figure 10-3 illustrates a D/A analog waveform.

The "stair-step" waveform contains significant high frequency energy associated with the sharp edge transitions. As with the A/D, a lowpass filter is required; here, it is used to *reconstruct* the analog waveform. These lowpass filters remove the sharp edges and restore the smooth analog waveform.

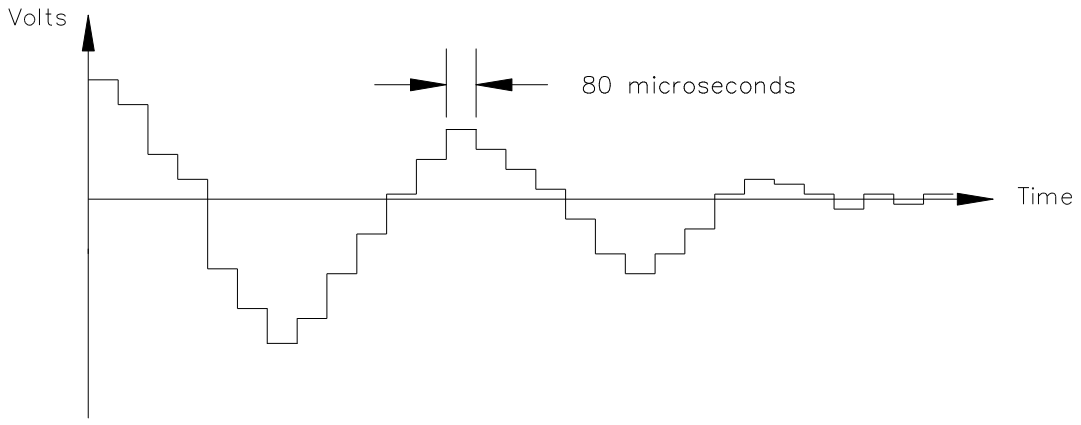


Figure 10-3: D/A Output Waveform

Like the sigma-delta A/D, the complementary *delta-sigma*, or "1-bit," D/A can further reduce reconstruction lowpass filter requirements. Samples at a rate of 2 F_N are passed to a special *interpolation* and *noise-shaping* digital filter chip and are converted to a measured rate of 64 F_N.

These high-speed samples are digital-to-analog converted and then lowpass filtered using a filter with a less sharp transition band, again typically implemented with a simple RC pair.

10.2 Digital Filters

The two standard classes of digital filters are *finite impulse response* (FIR) and *infinite impulse response* (IIR) filters. IIR filters have limited application in forensic voice processing and are not discussed here. The preponderance of applications use FIR filters.

Within FIR filters, two principal types are used in voice applications: *adjustable filters* and *self-tuning filters*. Adjustable filters have their control parameters (cutoff frequency, slope, stopband, etc.) adjusted by the operator. Self-tuning filters, which include adaptive and spectral inverse digital filters, analyze the audio signal and adjust their filters automatically to reduce noise.

10.2.1 FIR Filter

The FIR filter is functionally illustrated in Figure 10-4. This filter is constructed with delays, multipliers, and adders, and is usually implemented in a specialized digital computer employing multiple microprocessors. DAC instruments typically incorporate 20 or more DSP microprocessors to implement such filters.

The input audio from the analog-to-digital converter consists of a sequence of samples

$$x_1, x_2, x_3, \ldots, x_n, x_{n+1}, x_{n+2}, \ldots,$$

where x_n is the n^{th} audio sample. The corresponding output signal samples from the filter are

$$y_1, y_2, \ldots, y_n, y_{n+1}, \ldots.$$

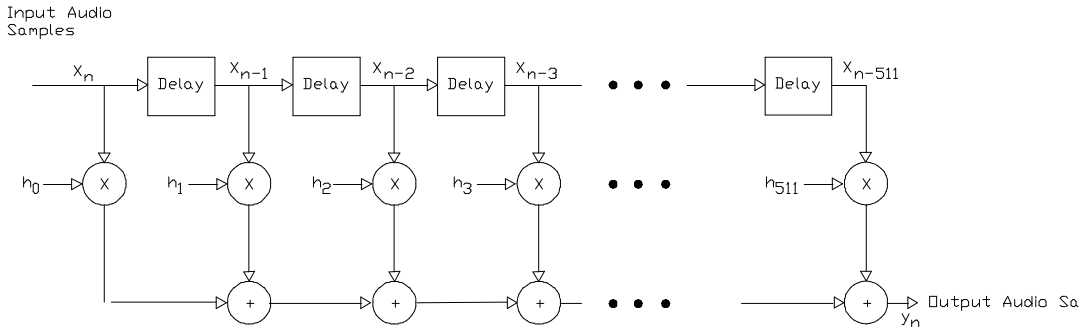


Figure 10-4: FIR Filter Structure

The FIR digital filter utilizes a *delay line*, which can be conceptualized as a first-in first-out queue (or "line") that has a fixed number of positions. Each time a new sample x_n is produced by the A/D, all samples are shifted one position, and the new A/D sample is placed in the now-empty first position. The sample at the output of the first delay box is now one sample interval old and is called x_{n-1}. In fact, all samples are shifted one step to the right. As a result, the filter always contains the most recent N samples (N is 512 in Figure 10-3) from the A/D. After a sample has been shifted N times, it is discarded and no longer remains in the delay line.

At each clock pulse, each sample in the delay line is multiplied by its filter coefficient h_i. The h coefficients specify the characteristics of the filter and determine the filter type (lowpass, bandpass, etc.) and control parameters (cutoff frequency, stopband attenuation, etc.). The products resulting from the $N + 1$ multiplies are added together to form the output sample y_n. An output sample is therefore produced each time a new input sample x_n is introduced to the filter. Thus, the FIR filter equation is

$$y_n = h_0 \cdot x_n + h_1 \cdot x_{n-1} + h_2 \cdot x_{n-2} + \ldots + h_N \cdot x_{n-N}.$$

or, more compactly expressed as

$$y_n = \sum_{i=0}^{N} h_i \, x_{n-i}.$$

As can be seen above, the filter's output is the weighted sum of some number of previous input samples. More complex filters require more samples and coefficients, each pair of which is referred to as a *tap*. FIR filter equations can be easily programmed into the DSP's high-speed digital computer provided that it has sufficient "MIPS rate" and memory to implement the desired filter(s).

> Example: A digital filter has 5 taps and a coefficient set of $h_0 = h_1 = h_2 = h_4 = 0$, $h_3 = ½$. What is its output equation?
>
> $$y_n = 0 \cdot x_n + 0 \cdot x_{n-1} + 0 \cdot x_{n-2} + ½ \cdot x_{n-3} + 0 \cdot x_{n-4}$$
> $$= ½ \cdot x_{n-3}$$
>
> The output samples y_n are the input samples delayed by the sample intervals and reduced in value to half of the input sample x_{n-3}.

The filter's *coefficient set* (h_0, h_1, h_2 . . ., h_N) is called its *impulse response*. These coefficients completely determine the characteristics (transfer function) of the filter in a manner analogous to the way that a DNA genetic sequence completely determines the characteristics of a given plant or animal. For example, one coefficient set may produce a 1 kHz lowpass filter, while another set produces a 4 kHz highpass filter. In fact, virtually any filter shape may be produced by merely changing the numeric values of these coefficients. Since the coefficients are held in the

digital computer's memory locations, they may be easily changed, making the FIR filter a very flexible signal processor.

Two types of digital filters are commonly used for audio processing: manually-adjusted and self-tuning filters. The self-tuning filters include spectral inverse and adaptive filters; these are discussed in Section 10.3.

10.2.2 Adjustable Digital Filters

Adjustable filters are manually tuned by the operator to achieve the desired audio noise reduction/enhancement results. The operator adjusts one or more filter values (parameters). Examples of adjustable filters include lowpass, highpass, bandpass, bandstop, notch, comb filters, spectrum equalizers, and graphic filters.

As pointed out in the previous section, an FIR digital filter implemented in a high-speed digital computer is a very flexible device. By changing the coefficients (h_0, h_i, …, h_n) in the filter's memory, the filter can take on a wide variety of personalities (transfer functions). Figure 10-5 illustrates a functional block diagram of an adjustable digital filter.

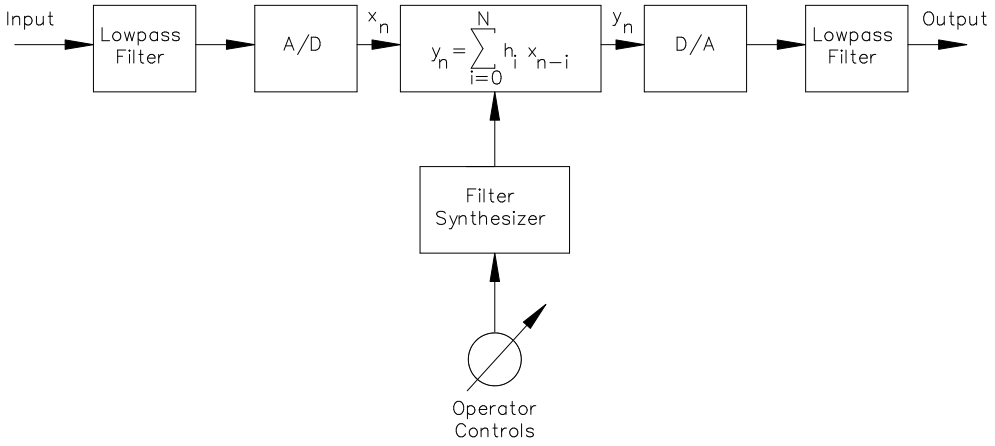


Figure 10-5: Adjustable Digital Filter

The input lowpass filter and analog-to-digital converter produce a stream of digital samples x_n from the input audio. These digital samples x_n are then passed to the digital computer. This computer implements one or more FIR filters whose stream of output samples y_n is passed to a digital-to-analog converter and lowpass filter, resulting in the processed output audio.

The high-speed digital computer is programmed with one or more FIR equations from Section 10.2.1, *i.e.*,

$$y_n = \sum_{i=0}^{N} h_i \, x_{n-i}.$$

The user sets the filter control parameters, such as the lowpass filter cutoff frequency and rolloff characteristics for each filter. The synthesis algorithm of the *filter synthesizer* programmed into the computer processes these user inputs to produce the filter's coefficient set h_0, h_1, h_2, . . ., h_N. This algorithm usually consists of a set of specialized equations and procedures for calculating these coefficients. Alternatively, the coefficient set may be derived in advance and stored in memory for recall by the user.

> Example: How many computations (multiplies, adds, and delays) are required for a 1024th order digital filter to process one output sample y_n? If the sampling rate is 40 kHz, how many computations per second are required?
>
> Each filter tap requires three operations (one multiply, one add, and one delay); 1024 taps require 3 × 1024 = 3072 operations per output sample.
>
> With a sampling rate of 40,000 samples per second, the number of operations required is
>
> operations per second = 40,000 Hz × 3072 operations
> = 122,880,000 operations/sec !

In order for a digital filter to be effective, it must operate in *real-time*. Real-time means that one audio sample is output every time one is input. Early DSP systems were not real-time. They would first sample the audio and store it on hard disk or digital tape in a general purpose computer. The digital computation would next be carried out producing filtered audio samples stored back on the disk or tape. The final step would convert these output audio samples back to analog form. The digital computation would often take many times real-time to execute in a general-purpose computer. A digital filter can easily require over 100,000,000 mathematical operations per second of audio. Often a 30-minute audio recording would take 100 or more hours to process.

DAC digital filters all operate in real-time. The operator can make adjustments to the filter and immediately observe the effects without waiting long periods of time.

10.2.2.1 Bandlimiting Filters

The FIR filter can implement conventional lowpass, highpass, bandpass, and bandstop filters. Such digital filters have substantial advantages over their analog filter counterparts. Digital filters may be implemented with linear phase characteristics (all frequency components have the same delay, hence no phase distortion), and very sharp cutoffs (many times sharper than analog

filters). FIR filters have finite memories (after *N* delays the input samples are discarded) resulting in substantially reduced ringing with impulsive noises.

Analog filters have proportional rolloff characteristics, expressed as dB/octave. This means that a filter's rolloff becomes less sharp the higher the frequency. Digital filters have constant rolloff characteristics, expressed in dB/Hz. The rolloff rate is unaffected by the cutoff frequency whether it is high or low. Digital filter rolloffs can therefore be much sharper than their analog counterparts. Slopes equivalent to 10,000 dB/octave are readily achievable, whereas analog filters are often limited to 100 dB/octave.

A frequency transfer curve for a digital lowpass filter is given in Figure 10-6. The frequency regions are partitioned into passband, transition band, and stopband. DAC digital filters allow *independent adjustment* of the cutoff frequency, transition slope (rolloff), and stopband attenuation. This type of bandlimiting filter is referred to as a *parametric filter*. Figure 10-7 illustrates the PCAP's lowpass filter control window.

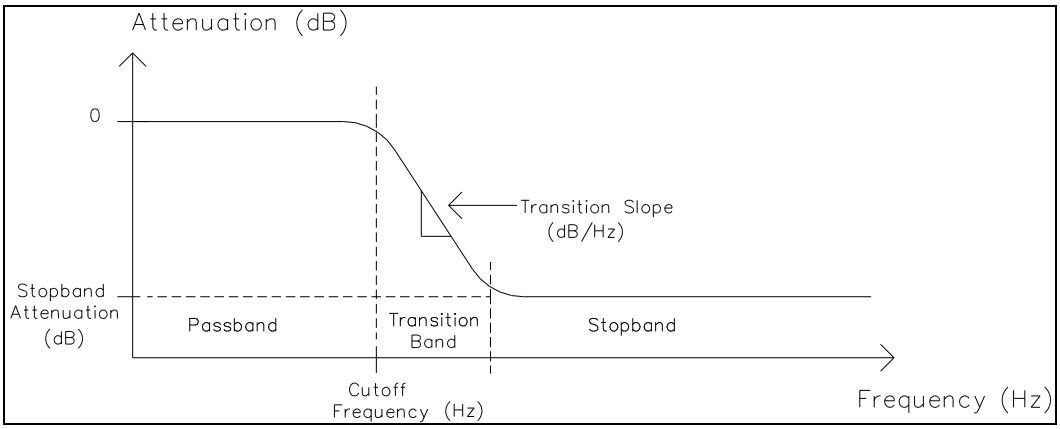


Figure 10-6: Digital Lowpass Filter Illustration

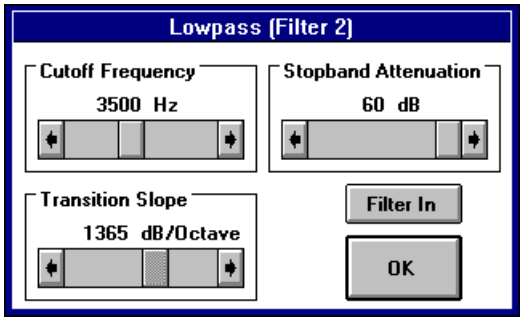


Figure 10-7: Lowpass Filter Control

10.2.2.2 Graphic Equalizers

Graphic equalizers were introduced in Section 9.2.3. The principal disadvantages of an analog graphic equalizer are not present in a digital graphic equalizer. The constant rolloff characteristics allow uniform narrow-band resolution for each slider frequency adjustment. Since a single, low-ringing FIR filter can be used, rather than 15 to 25 ringing analog filters, its response to impulsive noises is greatly improved.

DAC produces five digital graphic equalizers as part of its PC-based MCAP and PCAP systems and the ENHANCER and two Scrubber products.

10.2.3 Comb Filters

A special FIR filter used in removing *harmonic* noise, such as AC power hum, is a comb filter. AC hum is characterized on a spectrum analyzer as having strong energy concentrated at frequencies of 60, 120, 180, 240, ... Hz. A special multi-notch filter having notches at 60 Hz and multiples of 60 Hz (*i.e.*, harmonics of 60 Hz) is called a 60 Hz comb filter. This filter receives its name from its transfer function, which resembles a hair comb. Figure 10-8 illustrates a 60 Hz comb filter's transfer function. Note that notches occur at 60, 120, 180, . . . Hz.

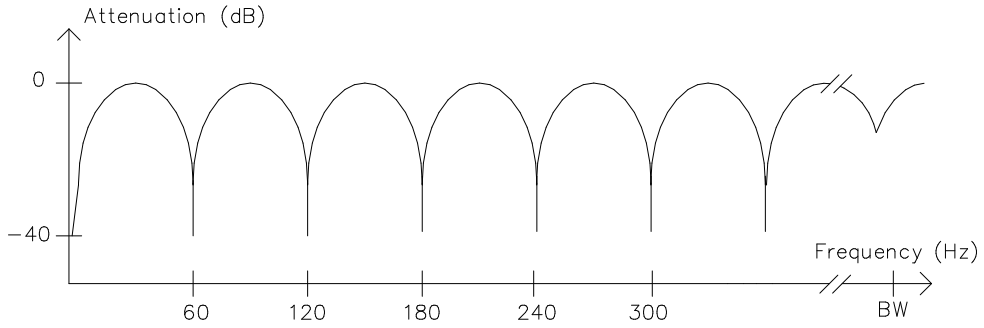


Figure 10-8: Transfer Function of a 60 Hz Comb Filter

Functionally, a comb filter is the simplest of digital filters. Figure 10-9 illustrates this device.

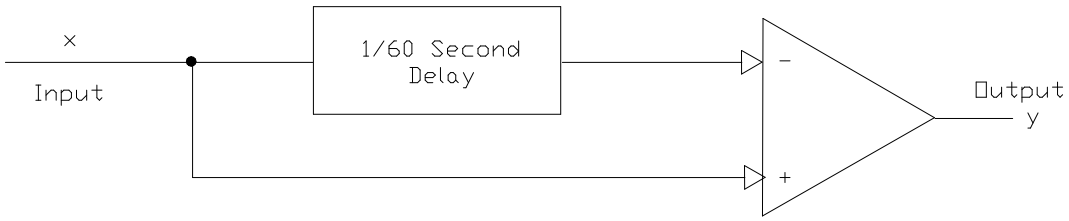


Figure 10-9: Functional Block Diagram of a 60 Hz Comb Filter

A sine wave is a *periodic* waveform, *i.e.*, one cycle is identical to the next. If a sine wave is delayed exactly one cycle (16.7 msec for 60 Hz) and subtracted from the original, undelayed sine wave, perfect cancellation takes place. This process also cancels harmonics of the sine wave, since the second harmonic is being delayed exactly two cycles; the third harmonic is being delayed exactly three cycles; etc.

A comb filter's fundamental frequency notch is determined by the amount of delay. If the delay were 1/50 sec = 20 msec, then the comb filter would have notches at 50, 100, 150, . . . Hz.

Typically, AC hum is very intense at low frequencies but insignificant at higher frequencies. Since a comb filter will produce notches all the way across the audio passband, a substantial amount of higher frequency energy may be notched out unnecessarily. Notch filters, like that shown in Figure 10-9, also introduce a mild reverberation. (The signal summed with itself delayed is a type of echo.) An improved comb filter is one which allows restricting the notches to a limit, for example, 1000 Hz. Consider the modified comb filter of Figure 10-10.

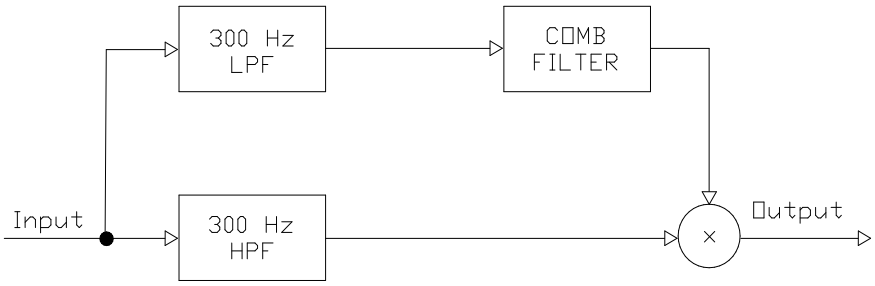


Figure 10-10: Notch-Limited Comb Filter

The input audio is split into two bands. The audio below 300 Hz is comb filtered, and that above 300 Hz is not. The two signals are then recombined to form a transfer function illustrated in Figure 10-11.

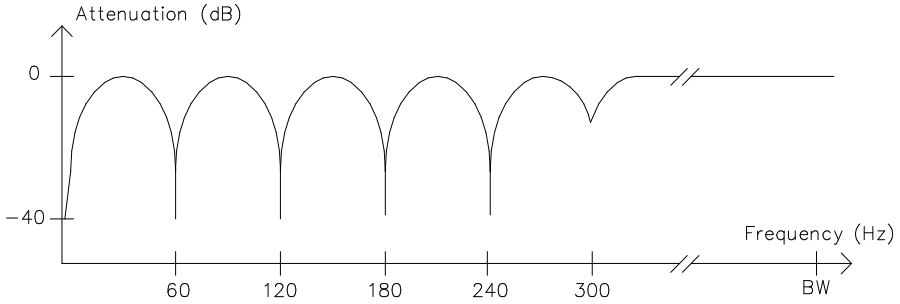


Figure 10-11: Transfer Function of Notch-Limited Comb Filter

In some situations it is desirable to control the depth. This is done by modifying the basic comb structure to allow the operator to specify the depth. Figure 10-12 illustrates the PCAP's comb filter control window.

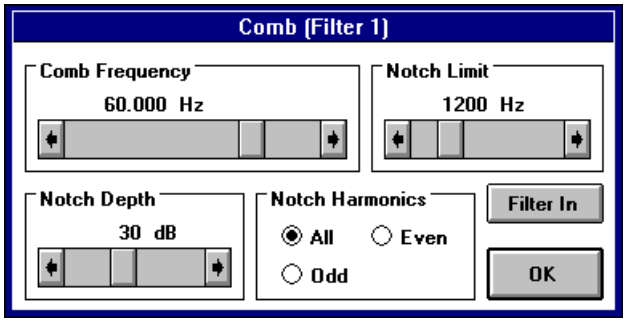


Figure 10-12: Comb Filter Control

The fundamental comb frequency, *e.g.*, 60 Hz, is normally adjusted by the operator to achieve maximum hum cancellation. Even though the original hum was at 60 Hz, a tape recorder may shift this frequency due to errors in tape speed. Comb filters allow adjustment of this frequency over a wide range, typically 40 to 400 Hz.

Comb filters are often able to set this fundamental frequency automatically by using a tracking circuit. Such circuits are sometimes not able to adjust the filter frequency as well as the operator can manually due to noise and other signal anomalies.

> Example: A cassette has a major hum component at 185 Hz. Since this cassette was recorded in North America (60 Hz AC power), how much faster is the player than the original cassette recorder?
>
> Answer: The third harmonic of 60 Hz is 180 Hz, which is the strongest component present.
>
> The player speed is, therefore
>
> $\frac{185\ \text{Hz}}{180\ \text{Hz}} = 1.028$, *i.e.*, 2.8 percent faster than the recorder.

From the above example, reducing the playback speed by 2.8 percent would match the playback speed to that of the original recorder. This has two advantages. First, all audio frequencies would be properly reproduced (though the ears would probably not detect a modest 2.8 percent frequency shift). Second, the comb filter would be set to a 60 Hz notch frequency and the variable-speed tape player would be easily for maximum hum removal.

Another issue illustrated by the example is that the fundamental (60 Hz) and second harmonic (120 Hz) are much reduced in energy. Normally during recording, 60 Hz is the loudest

component; however, many recorders will not pass 60 Hz very well and will attenuate such low frequencies.

One would think that 120 Hz, which is often recorded well, would be the strongest *recorded* component; however, this is rarely the case. Due to the symmetry of the AC waveform, the odd harmonics (180, 300, 420 Hz, etc.) are often much stronger than the even harmonics (120, 240, 360, etc.).

To allow the forensic examiner to exploit this, the comb filter control window of Figure 10-12 permits notching only odd or even harmonics as well as both odd and even (all). In this manner, the minimum number of notches can be utilized.

As mentioned earlier, a comb filter introduces a mild echo which can be minimized by limiting the number and depth of notches. It is recommended that a one-channel adaptive filter be used to reduce the echo and to attack components missed by the comb filter. Analog recordings are subject to wow and flutter, which causes the hum frequency to wobble in and out of the notches. The adaptive filter will assist in cleaning up this residual hum energy.

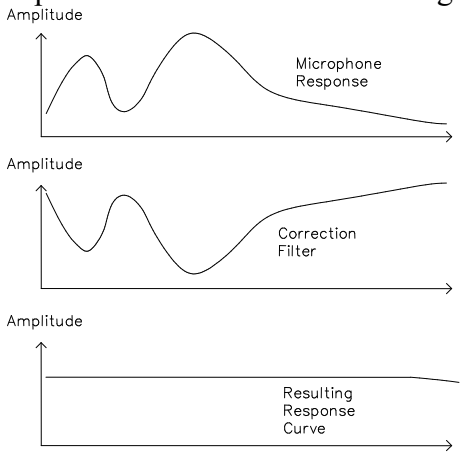


Figure 10-13: Graphic Filter Correction of Microphone Response

10.2.4 Graphic Filters

Graphic filters allow the user to "draw," usually with a mouse, the filter shape (called its transfer function). This type of filter is very useful in equalizing a signal and in compensating for microphone and concealment deficiencies. Suppose a microphone is measured to have the response curve shown in Figure 10-13. Microphone response curves are usually supplied by the manufacturer. These curves assume no influence of nearby acoustic materials such as the mounting or concealment fixture. By precisely "drawing" an inverse curve with the mouse, an equalizing filter can be created.

The microphone signal can thus be "corrected" by passing it through the graphic filter prior to recording. See Figure 10-14.

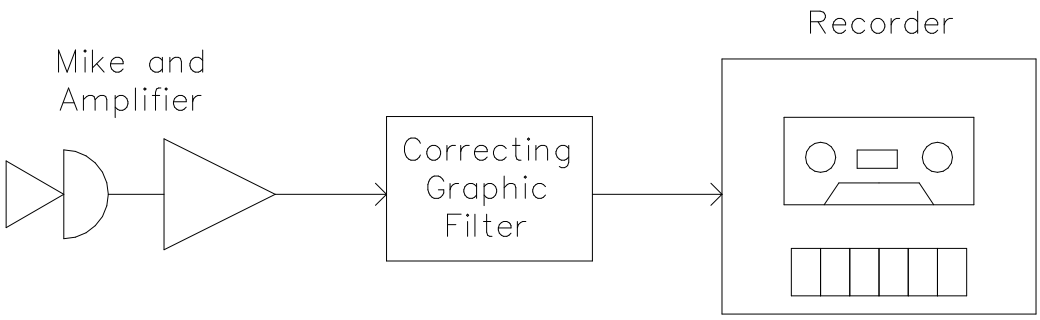


Figure 10-14: Microphone Equalization Setup

Graphic filters may be used to replace parametric and graphic equalizers. A graphic filter is characterized by two parameters: *resolution* and *dynamic* range. Resolution is the precision with which frequency can be specified. Figure 10-15 illustrates the graphic filter in the PCAP. This filter has 460 lines of resolution spread across (in this example) 6.5 kHz. The frequency resolution is thus

$$\frac{6500}{460}\ \text{Hz} \;=\; 14\ \text{Hz}$$

The dynamic range specifies the maximum signal attenuation range. In the figure, the dynamic range is 60 dB, which is more than adequate for virtually all forensic audio applications.

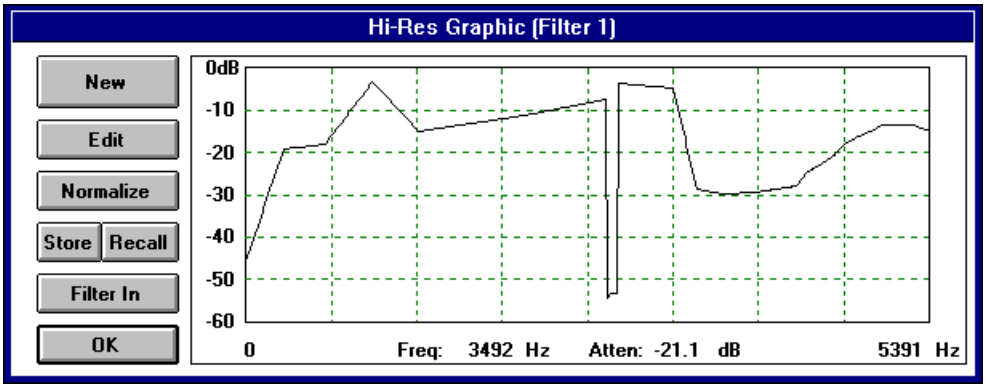


Figure 10-15: PCAP Graphic Filter Display

Graphic filters are computer controlled and offer a number of operator conveniences. Filters may be stored and recalled from memory. The mouse cursor specifies the end point of each segment of the filter curve as it is drawn; frequency and attenuation coordinates are continuously displayed as the mouse is moved. Also, the curve may be edited with the mouse; the region to be changed is specified, and a new curve is drawn in that region.

10.3 Self-Tuning Digital Filters

Self-tuning digital filters automatically adjust themselves to obtain noise reduction on audio signals. Such filters have special algorithms that make noise measurements on a signal and automatically calculate the FIR filter coefficients h_0, h_1, h_2, . . ., h_N. The actual signal processing is carried out in an FIR filter.

Self-tuning filters include spectral inverse and one-channel adaptive filters. A spectral inverse filter (SIF) is a spectral flattening (whitening) filter which attempts to smooth out irregularities (usually caused by noises and resonances) in the audio power spectrum. A deconvolution filter attempts to achieve the same goal using a more precise statistical analysis of the signal. The principal deconvolution filter used in forensic audio is the one-channel adaptive filter.

A self-tuning filter consists of an FIR digital filter for filtering the audio and a filter design (or "synthesis") algorithm for adjusting the FIR filter. Figure 10-16 illustrates the process. In order to properly adjust the filter, the synthesis algorithm must observe the signal before filtering (called "feedforward") or after filtering (called "feedback") or both.

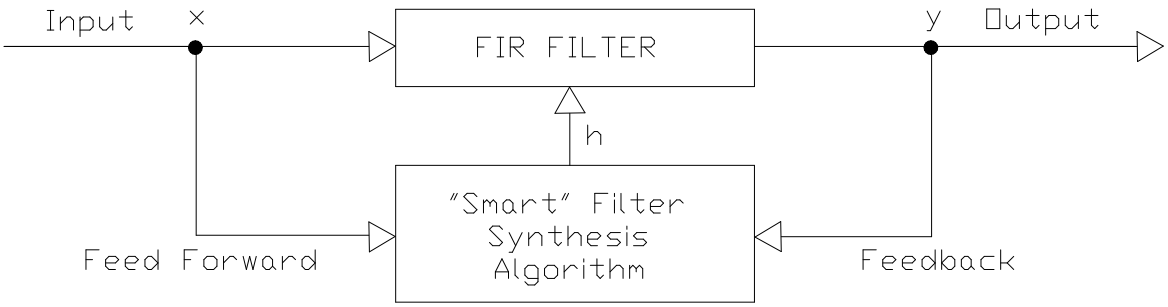


Figure 10-16: Self-Tuning Digital Filter

10.3.1 Spectral Inverse Filter (SIF)

Spectral Inverse Filtering (SIF) is a powerful automatic equalization procedure that attacks muffling, resonances, acoustic effects, and tonal noises on voice audio. This procedure works best when the noises are *stationary*, *i.e.*, their characteristics do not change. This filter cannot track and remove music; the adaptive filter of Section 10.3.2 is better suited for that task.

The SIF has been found especially useful with recordings having significant wow and flutter. The SIF responds to average noise effects and can often better cope with tape recorder speed variation.

A spectral inverse filter flattens the audio spectrum by applying a special filter to the audio. This special filter has a shape which is *opposite*, or inverse, to the power spectrum.

Consider Figure 10-17. The original audio spectrum is given at the top; the spectral inverse filter shape is given in the middle; and the resultant audio spectrum is given at the bottom of the figure.

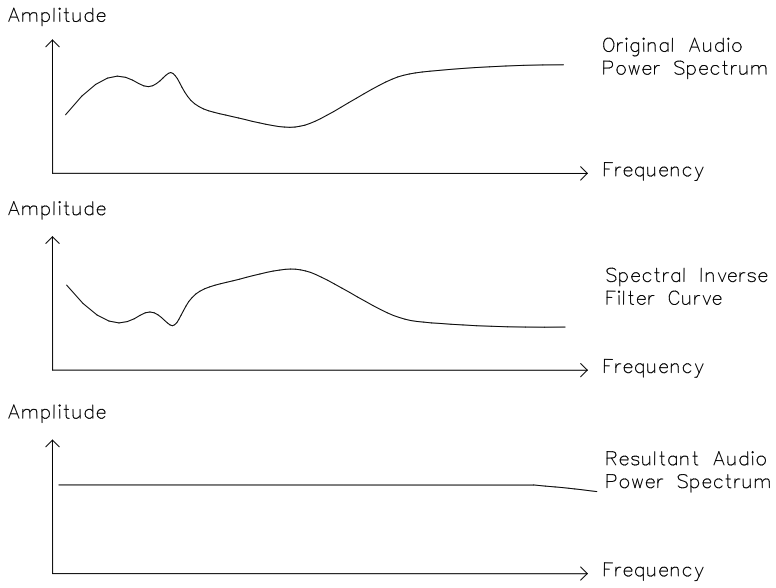


Figure 10-17: Spectral Inverse Filter Illustration

A functional block diagram of a spectral inverse filter is given in Figure 10-18. The input audio signal's power spectrum is measured with an FFT spectral analyzer. The PCAP and MCAP have the FFT analyzer built into the system. The measured power spectrum is averaged over several seconds to give *long-term spectral* information. Short-term spectra change rapidly and are more influenced by voice changes. Long-term spectral information smoothes out the effects of the voice and yields information about acoustic resonances and stable noises (such as tones and hum).

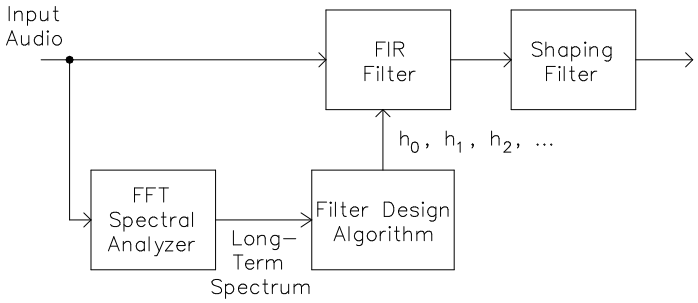


Figure 10-18: SIF Functional Block Diagram

The Filter Design Algorithm designs a set of *h* coefficients for the FIR filter from the long-term power spectrum. The resultant digital filter has a shape that is opposite that of the input signal's spectrum. The filter design procedure is akin to the FFT spectral analysis procedure and employs an FFT and window. The process is called *frequency sampling filter synthesis* and results in a

linear phase (no phase distortion) digital filter. The actual synthesis procedure is beyond the scope of this document.

The FIR filter's output audio has a flat audio spectrum which is somewhat unnatural sounding. Voice typically has a high frequency rolloff. See the voice long-term power spectrum in Figure 2-3. An output shaping filter can "reshape" the flat spectrum to one that sounds more like a normal voice. On DAC filters, several desired output shapes are selectable. See Figure 10-19 for the PCAP's SIF control screen.

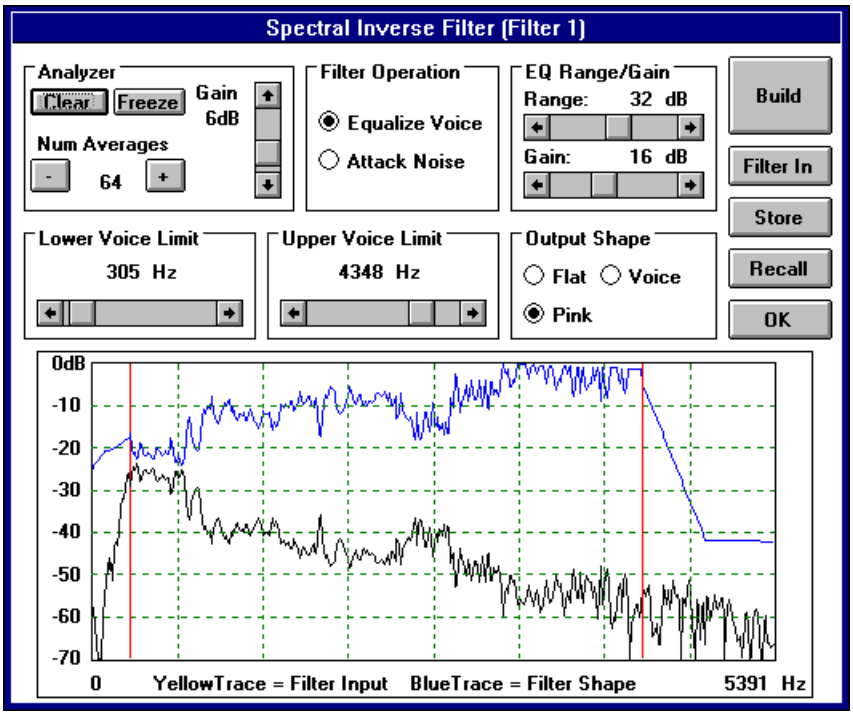


Figure 10-19: PCAP Spectral Inverse Filter Control Screen

Like the Graphic Filter, the SIF is characterized by both resolution and equalization (dynamic) range. The PCAP's SIF has 460 lines of resolution, resulting in 14 Hz frequency resolution at a bandwidth of 6.5 kHz (6500 Hz ÷ 460 = 14 Hz). The PCAP's SIF equalization range is 50 dB, which means that it can flatten spectra that have variations of as much as 50 dB.

The equalization range is controlled by the operator and specifies the maximum amount of *attenuation* that is applied to the spectral peaks. If the equalization's range is set to maximum, the audio will be completely flattened. Smaller equalization ranges will result in only the spectral peaks being flattened. The spectral valleys will retain their original shape.

Since the flattening involves "shaving" off the peaks (via linear attenuation), amplification, or gain, must be applied to restore the audio level. Figure 10-19 has both EQ (equalization) Range and Gain controls.

When using the SIF to equalize the voice, an upper and lower voice frequency limit are specified. These are the lower and upper sounds where voice energy is present on the audio signal, *e.g.*, 250 and 3750 Hz. Between these two limits, the audio (voice) is flattened; outside these limits, the audio is rolled off, or bandpass filtered. This process is very effective where severe resonances are present and affect voice quality.

The SIF may also be used to attack specific bands of noise. Suppose an automobile horn occurs in the range of 500 to 1100 Hz. By setting the Lower and Upper limits to 500 and 1100 Hz and selecting the Attack Noise mode, only that 600 Hz-wide band is equalized. The frequencies below 500 Hz and above 1100 Hz are passed on unaffected.

Examples of Spectral Inverse Filters:

In the examples following, SIF will equalize a voice spectrum using various **EQ Ranges** and **Output Shapes**. When the Range is small, only the peaks in the spectrum are flattened. As the Range is increased, lower energy segments are equalized. The top trace in each of the figures below gives the filter curve and the bottom trace gives the original input spectrum. In Figure 10-20, Figure 10-21, and Figure 10-22, the EQ Range is increased. Each increase in Range is accompanied by an increase in compensating Gain. Compare the filter curve (top) to the original input spectrum (bottom). As the **EQ Range** is increased, more peaks are attenuated. Figure 10-22 completely compensates all peaks and valleys. Note also the 40 dB attenuation outside the two Limits.

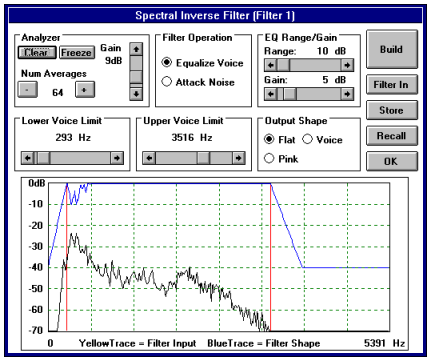


Figure 10-20: EQ Voice Operation, EQ Range Set to 10dB, Output Shape Set to Flat

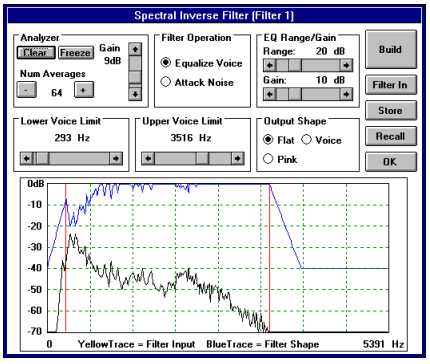


Figure 10-21: SIF with EQ Range Set to 20dB

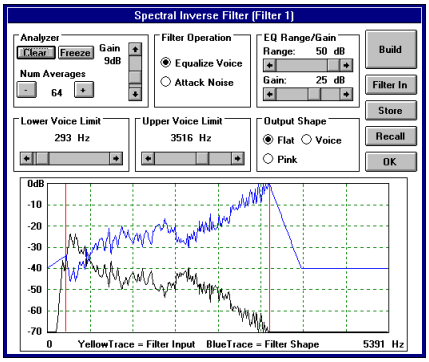


Figure 10-22: SIF with EQ Range Set to 50dB

The Output spectral Shape may be selected as Flat, illustrated in Figure 10-20, Figure 10-21, and Figure 10-22 above. It may also be set to Voice or Pink. See Figure 10-24.

The Attack Noise mode does not attenuate Out-of-Limits signal, but equalizes and attenuates in-Limits signal frequencies. Figure 10-23 illustrates.

SIF Applications Suggestions

The following suggestions may be beneficial in setting up and operating the spectral inverse filter.

Analyzer: The FFT spectrum analyzer automatically produces the power spectrum of the signal entering the SIF. For the SIF to be effective, a smoothed spectrum is necessary; SIF adjusts for long-term stable noises including resonances and steady noises. Short-term effects of voice and nonstationary noises will decrease the filter's effectiveness; therefore, the **Num Averages** parameter should be set large. At least 32 averages are recommended, but more will give a smoother spectrum from which to build a filter.

The Analyzer **Gain** should be increased to display weaker energy components. Do not overload (OVL) the analyzer, as the spectral information would become corrupted.

Try to capture a representative spectrum using the Freeze/Run button. Once a stable, representative spectrum is obtained on the display, Freeze the analyzer. This same spectrum may be used for several different variations of the filter (changing Limits or Range, as an example).

NOTE: Immediately following Freeze, the screen will continue to update briefly, as the PC is receiving the final pre-frozen spectral data from the external unit.

Limits: In the Equalize Voice Operation, the Upper and Lower Voice Limits should bracket the voice signal. Outside these limits, the audio is bandpass-filtered. Setting the Lower Limit above 300 Hz and the Upper Limit below 3000 Hz may adversely affect intelligibility. Try several sets of Limits; build the SIFs; Store the filters, and Recall and compare in an A/B fashion.

In the Attack Noise Operation, bracket the noise band surgically with these Limits. If there are several disjoint noise bands, series additional SIFs. (The PCAP will series up to four). The audio outside those limits is unfiltered.

Range and **Gain**: SIF is a spectral attenuator. The spectral peaks are pressed down toward the spectral valleys. The amount of reduction is limited by the **Range** scroll bar. If, however, all peaks are reduced to the *lowest* valley, no more reduction takes place.

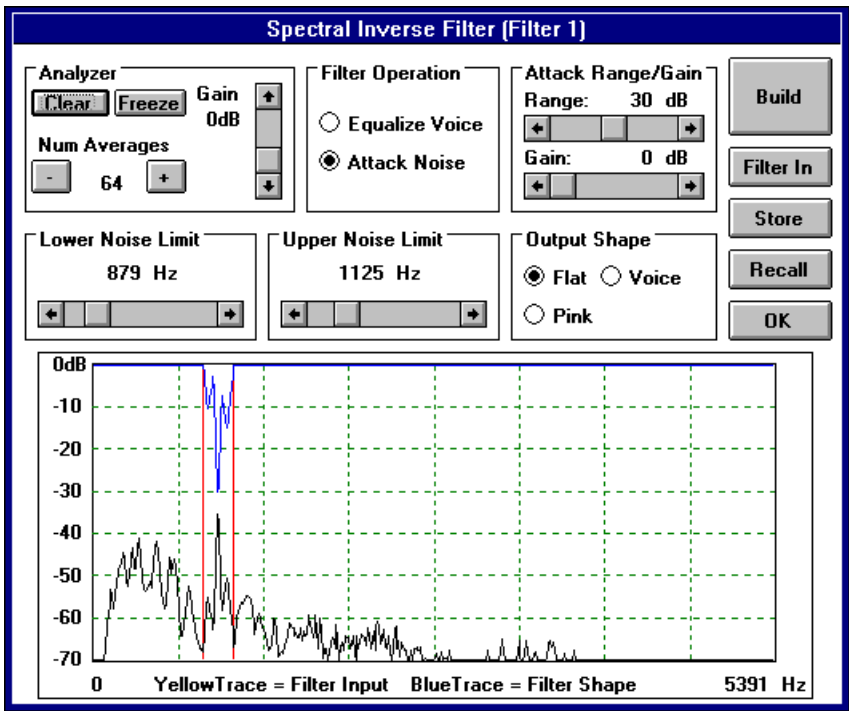


Figure 10-23: Attack Noise Operation, Attack Range Set to 30dB

For example, if the distance from the highest peak to the lowest valley *between the Upper and Lower Limits* is 30dB, the maximum range actually used will be 30dB, even if the scroll bar is set to a larger **Range**. Normally a **Range** of 10-20dB is adequate, as attacking the stronger spectral peaks will provide the necessary enhancement. Go gently at first and compare several SIFs with different **Ranges**. The Store and Recall memories are useful in saving candidate filter solutions.

Since a SIF is an attenuator, the audio level should be restored with the **Gain** scroll bar. Normally, setting this to one-half the **Range** is adequate. If output distortion occurs, reduce the **Gain**. The **Gain** may be increased to make up for losses in previous Filter stages. Note that the Gain is usually 0 dB in the **Attack Noise** mode, but often some boost needs to be applied to elevate the level of the voices after the strong noises have been removed.

Output Shape: Three output spectral shapes are selectable by the user. The **Flat** shape requires the SIF to produce a uniform (to the degree possible) long-term output spectrum. This can be subsequently *reshaped* by the Hi-Res Graphic Filter or Output Equalizer to a more natural-sounding spectrum. Alternately, the **Voice** or **Pink** shapes may be selected, each of which has built-in output shaping. *The **Pink** shape is often the most pleasing to the ear.*

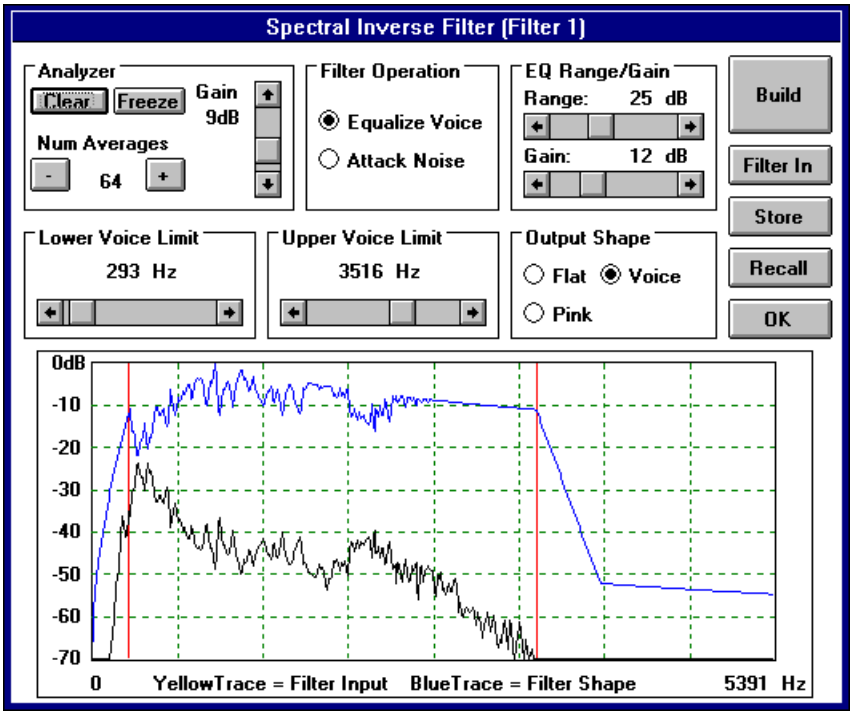


Figure 10-24: SIF with Output Shape Set to Voice

Experiment: Select a section of audio and display its spectrum. *Freeze* the analyzer and vary different control settings, storing built filters. Compare the results and determine the best solution. Always compare these different filter solutions using the same input audio.

A **Store** and **Recall** capability is also provided to allow the user to store commonly-used filter shapes to disk memory so that they can be recalled for later use. This capability also allows the comparison of candidate solutions during an enhancement session.

10.3.2 One-Channel Adaptive Filter

The one-channel (1CH) adaptive filter seeks to reduce noise by eliminating time-correlated events in a signal. This can be an extremely powerful tool because speech is principally random (not time-correlated) in nature, while noises are often highly correlated.

Consider the illustration of Figure 10-25. Here the random voice is added to correlated noise in the room. The resulting microphone signal is the sum of these two components.

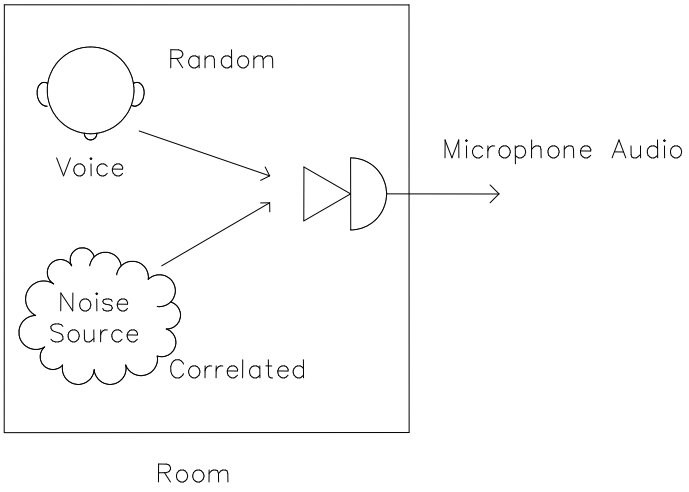


Figure 10-25: Correlated and Random Audio Combination

10.3.2.1 Predictive Deconvolution

Predictive deconvolution is a powerful noise cancellation procedure which *decorrelates* the input signal, removing long-term correlated signal components (*i.e.*, undesired noises). This process separates time-correlated and random signal components using a signal *predictor*. Time-correlated components are highly predictable and therefore readily separated from the remaining signal.

Consider the sine wave (tone) of Figure 10-26. The voltage fluctuates from +10 volts to –10 volts over time. If the clock is frozen at the present time A, then the voltage at future time B can easily be *predicted*. Merely look at the past; observe the repeated cycles, and estimate the value in future. A tone is a very predictable waveform.

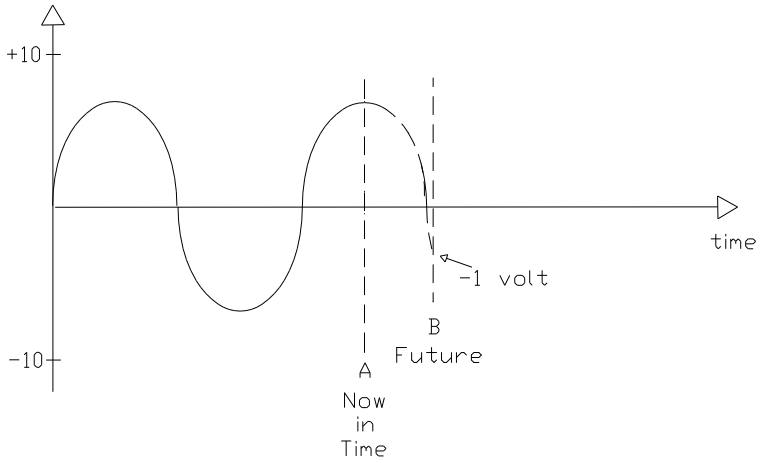


Figure 10-26: Predicting a Tone Wave

Many additive noises are also similarly predictable. Tones, music notes, and hum are all strongly time-correlated in nature. Because music changes from note to note, a *tracking* predictor, *i.e.*, an adaptive filter, is required to attack music effectively. Reverberations and echoes are also very predictable, *once the original sound occurs*. Though the original sound itself may not be predictable, the room's response to that sound is once we have observed it.

The signal predictor thus analyzes the recent history of the waveform and estimates what will occur next. If the waveform is strongly correlated, then the estimate will be very accurate. If the waveform is weakly correlated, then the estimate will be less accurate. If the waveform is completely random, however, the predictor will estimate zero (no estimate possible). Thus, predictive deconvolution is completely ineffective for reducing random noises such as RF static.

Figure 10-27 illustrates the process. The input audio signal is sampled yielding the sequence x_n, $n = 1, 2, ...$ A linear transversal predictor is an FIR digital filter and estimates the next signal sample x_n, yielding an output residue sample:

$$e_n = x_n - p_n. \qquad (1)$$

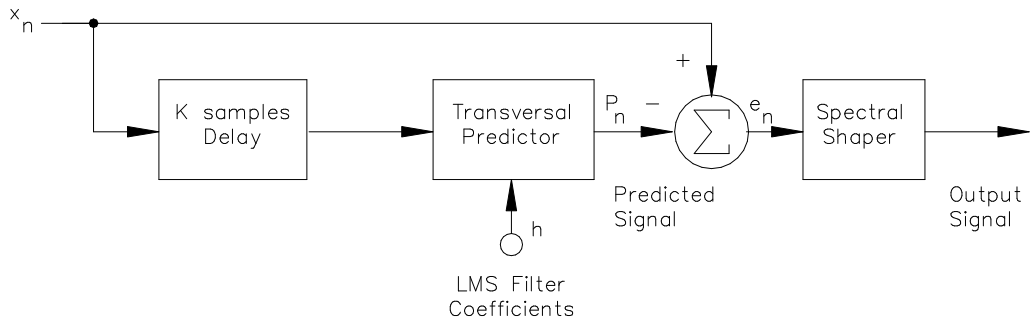


Figure 10-27: Predictive Deconvolution

Because a least-mean-square (LMS) estimator is used, the output residue signal is decorrelated in a long-term sense and generally has a flat power spectrum. A spectral shaper is thus often needed to restore the long-term output power spectrum to the desired shape.

The predictor used in this procedure is an FIR digital filter whose output estimate p_n is the weighted sum of past input signal samples, *i.e.*,

$$p_n = \sum_{i=K}^{K+N} h_i x_{n-i}, \qquad (2)$$

where each h_i ($i = K, K + 1, ..., K + N$) is a filter coefficient. K is the *prediction span*.

The filter order or size N required for the predictor depends upon the complexity of the time-correlated signal components being removed. Simple tones added to voice signals can be removed with N being small, *e.g.*, less than 20. Reverberation, however, generally requires a much longer filter. Acoustic reverberations of rooms generally require filter sizes greater than 1024 to adequately model room acoustics and longer echo paths.

The prediction span K is the time into the future that the signal is predicted. This delay term specifies the degree of time correlation required for the predictive deconvolver. Few signal components are strictly predictable (time-correlated) or unpredictable (random). Different types of signals have varying degrees of predictability. Speech, for example, is mostly unpredictable; it can be predicted only a short time into the future, K, and then some components could be lost. Hum and reverberations are very steady (stationary), predictable signal components and are predictable over a large range of Ks.

By adjusting the K to be too large to attack the voice but still small enough to reduce the quasi-predictable noise, the predictive deconvolver is useful in filtering voice.

Derivation of the predictor filter coefficients h_i requires the computation of *a priori* signal statistics and the solving of a large set of N simultaneous equations. Though predictive deconvolution provides an optimal noise canceling process, it also contains two inherent disadvantages. First, the estimation of the *a priori* autocorrelation statistics and the solving of the N simultaneous equations require a significant amount of computing, especially when N is large. Second, the coefficient solutions are based on the assumption of long-term stationarity of the noise components, which often is not the case. These two disadvantages can be avoided and hardware realization more easily achieved by making the process adaptive.

10.3.2.2 Adaptive Predictive Deconvolution (One-Channel Adaptive Filtering)

Adaptive predictive deconvolution, illustrated in Figure 10-28, is achieved by using an adaptive processor to continually analyze both the input and output signals and to automatically derive the predictor filter coefficients h_i.

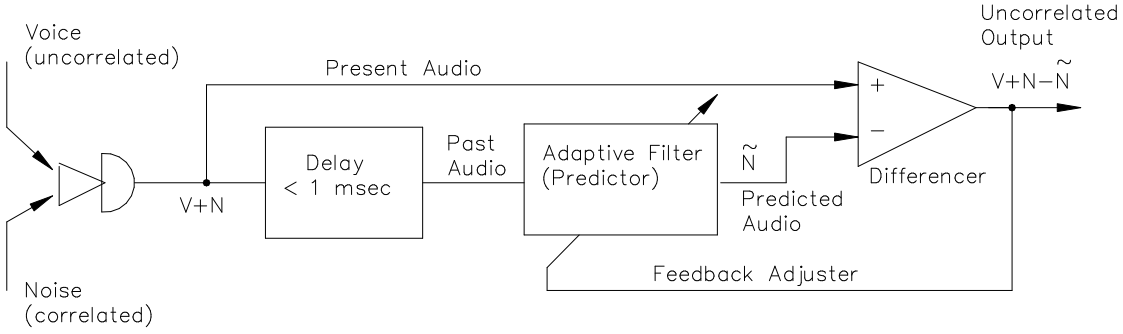


Figure 10-28: Adaptive Predictive Deconvolution (1CH Adaptive Filter)

The filter coefficients are derived using the least-mean-square (LMS) algorithm. The coefficients, which are initially all zero, are updated each sample interval according to the following procedure:

$$h_n(i + 1) = h_n(i) + \Delta h_n(i). \qquad (3)$$

Here $h_n(i)$ is the n^{th} predictor coefficient at the i^{th} time sample.

That coefficient is incremented by $\Delta h_n(i)$, which is computed using the LMS algorithm in the following manner:

$$\Delta h_n(i) = \mu \cdot e_i \cdot x_{i-n} \qquad \text{(LMS) (4)}$$

The parameter μ adjusts adaptation convergence rate. If μ is made large the filter will adjust itself in large steps, *i.e.*, converge very rapidly. If μ is set too large, the filter will over correct (too large values of Δh_is) and the coefficients will diverge. The result is a filter *crash* and results in severe output noise from the filter. Faster convergence rates (larger values of μ) are used for changing noise, such as music. Slower convergence rates (smaller values of μ) are used where the noise is more stationary, such as 60 Hz hum or reverberations. By setting $\mu = 0$, the filter is *frozen*, *i.e.*, $\Delta h_i = 0$ and the filter does not adjust itself. Freezing the adaptive filter still permits signal filtering to take place but inhibits the filter from readjusting itself to changing noises.

Some adaptive filters, such as those in the PCAP, have *conditional adaptation*. This feature enables or disables (freezes) adaptation based on the input or the output audio level's exceeding or falling below a user-adjusted threshold. Conditional adaptation may be used to freeze the filter when a loud voice is present (*i.e.*, freeze when the input level exceeds a large threshold), which allows the filter to adapt only when the background noise is present. Alternatively, the filter may be configured to adapt when strong noises are present (*i.e.*, adapt when the input level exceeds a small threshold) and freeze when the lower-level voice is present.

In addition to not requiring any prior knowledge of the signal and eliminating the need to solve the simultaneous equations, this adaptive approach allows time-variant noise sources to be canceled as long as these sources are correlated long enough for sufficient filter convergence. In practice, convergence times of a few hundred milliseconds are reasonable for N equal to a hundred or so. This convergence property allows, for example, the tracking and cancellation of music on voice signals or engine noises that vary as a function of changing RPM.

10.3.2.3 Adaptive Line Enhancement

Adaptive line enhancement is implemented in exactly the same manner as adaptive deconvolution (as shown previously in Figure 10-28) except that the predicted audio $\tilde{N}$ is taken as the output instead of the uncorrelated output $V + N - \tilde{N}$. This allows the predictable, time-correlated signal to be retained and enhanced, whereas the unpredictable random signal components are suppressed.

Normally, the adaptive line enhancer (ALE) has limited value in forensic audio enhancement, but it is useful in a few special cases. For example, if the signal of interest is not a voice but rather is a more predictable signal such as DTMF tones, modem signaling, or engine sound, then ALE can be used to reduce the random background signals and thereby aid the forensic analysis of the desired signal.

Another specialized use of ALE is for random noise suppression. As an example, consider the case of a covert recording made in a bathroom where a shower is running. In such a situation, the voices are relatively well correlated, at least on a short-term basis, whereas the shower noise is completely random. The intelligibility of the voices can sometimes be improved significantly by using the 1CH adaptive filter of the PCAP with a very small filter size, very large adaptation rate, short prediction span, and the predict output option.

10.3.2.4 Adaptive Filter Controls

Figure 10-29 shows the PCAP II's one-channel adaptive filter control window.

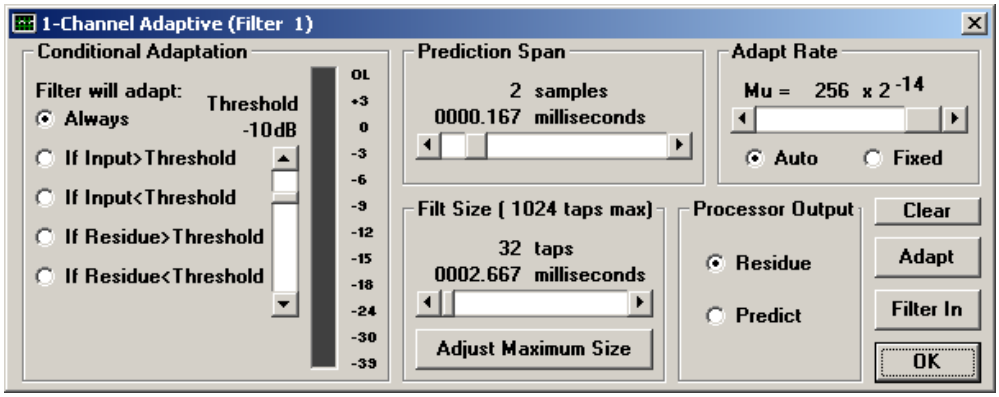


Figure 10-29: One-Channel Adaptive Filter Controls

A description of the controls of the 1CH adaptive filter follows:

Conditional Adaption: The adaptive filter can be forced to adapt only when the input (or output) audio exceeds (>) or falls below (<) an adjustable threshold.

Applications include:

1. Adapting only during sporadic talking (input > threshold), and
2. Not adapting during loud noise bursts (input < threshold).

Filter Size: This control sets the size of the digital filter. A 500 to 1000 tap filter meets most requirements. Shorter filters adjust more rapidly and are recommended for rapidly changing, less complex noise, such as music. Larger filters are required for complex noises such as raspy AC buzz. The filter memory time (filter size/sample rate) is also displayed.

Adapt Rate: This control specifies the speed at which the filter adjusts itself to changing noises. A small μ (mu) may be used for stable noises such as AC hum. A large μ is needed for dynamically changing noise such as music.

Prediction Span: This control sets the number of samples in the prediction span delay line. Prediction span is indicated both in samples and in

milliseconds, and it can be adjusted from 1 to 10 samples. Shorter prediction spans allow maximum noise removal, while longer spans preserve voice naturalness and quality. A prediction span of 2 or 3 samples is normally recommended.

Adapt Mode: Selects Automatic (Auto) or Fixed adaptation rate. Auto is recommended. When Fixed is selected, the specified Adapt Rate μ is applied to the filter at all times. However, when Auto is selected, the specified Adapt Rate is continuously rescaled (power normalized) depending upon the input signal level. Overall convergence rate is faster with Auto.

Processor Output: Selects Residue or Predict output signal. For adaptive predictive deconvolution, which is the normal mode of operation, use Residue output. For adaptive line enhancement, use Predict output.

Clear Button: Used to reset the coefficients of the 1CH Adaptive filter to zero without affecting any other adaptive filters in the process.

10.3.3 Spectral Subtractive Filtering

Spectral subtraction is a technique developed in the early 1970s for reducing broadband random noise on voice signals. This *nonlinear* filtering process is illustrated in Figure 10-30.

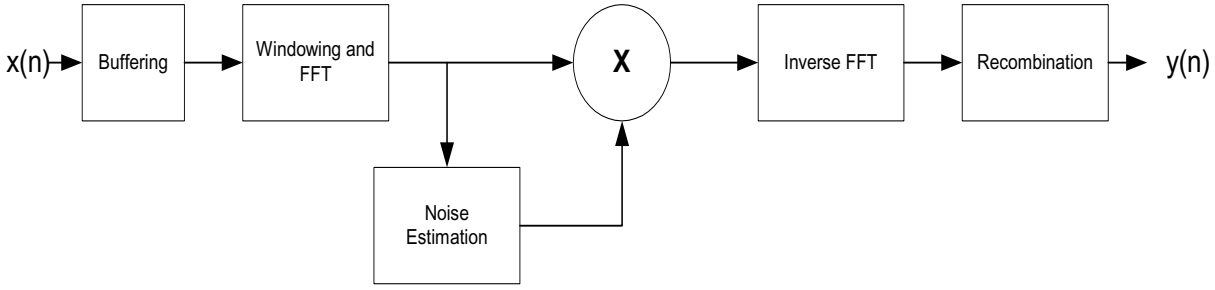


Figure 10-30: Simplified Spectral Subtraction Filtering

In the figure, the audio consisting of voice plus random noise is converted to the frequency domain using a Fourier transform. This transform produces spectral amplitude and phase components. A measured noise spectrum (amplitude only) is subtracted from the input audio's spectrum, hopefully removing the noise energy from the audio. This noise spectrum is derived from analyzing non-voice time segments of the audio. Because the resulting noise-subtracted amplitude spectrum may have negative values (negative energy), conditioning is applied to make all frequencies have at least zero energy. The original (unfiltered) phase and filtered amplitude components are then processed by an inverse Fourier transform for reconversion back to the time domain.

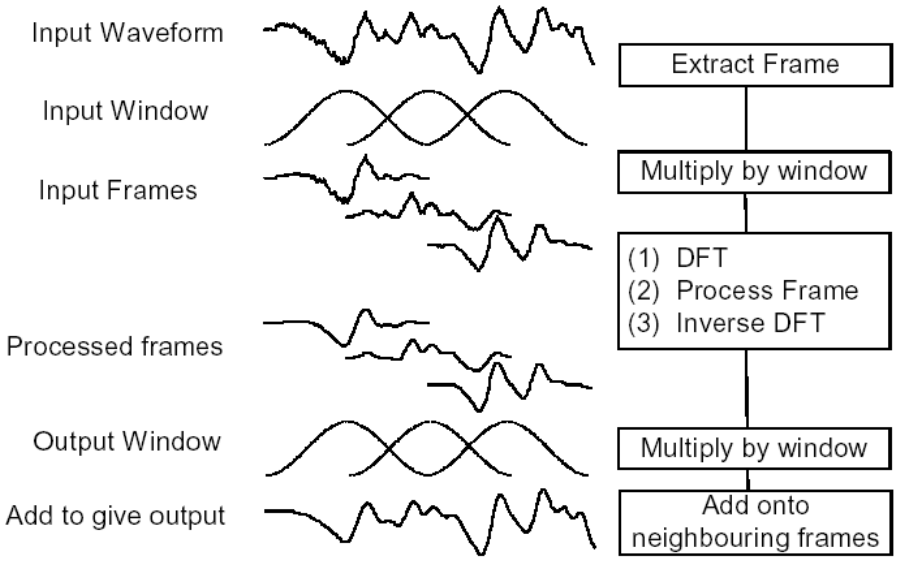


Figure 10-31: Illustration of Frequency-Domain Processing

This procedure, which is occasionally helpful in RF communications, has limited applicability to forensic audio enhancement. Since a digital Fourier transform operates on small intervals (typically 50 msec) of audio, the audio stream is segmented into blocks, or frames. Each audio frame is processed and concatenated at reconstruction, as shown in Figure 10-31.

Since the phase spectrum is unprocessed and remains noisy, the reconstructed audio stream consists of a sequence of discontinuous waveforms. To smooth out these popping sounds, the transforms are overlapped and the discontinuities are feathered out using a smoothing window. Still, “birdy noise” discontinuities often remain and the processed speech may be “mushy” and “gurgly”.

Often, the level of random noise is reduced by this processing, but the speech is nearly always adversely affected, actually reducing intelligibility. The main benefit of spectral subtraction is as a final “polishing” filter; once other non-random noises have been reduced by other filters, and the voices are otherwise intelligible, the spectral subtraction filter can be used to get rid of that last annoying bit of random noise, often with very good result. Care must be taken, however, to avoid introducing excessive audible artifact, otherwise the forensic validity of the enhanced recording may be called into question.

EXERCISES

1. An analog-to-digital converter samples at 12 kHz. What is the maximum audio frequency that can be sampled without aliasing distortion? What is the practical bandwidth of this setup?

2. What is the maximum dynamic range of a linear 8-bit A/D?

3. How much memory does a 2048^{th} order FIR filter have at a sampling rate of 10 kHz?

4. An audio signal has tones present at 240, 300, 360, and 480 Hz. To what frequency should the comb filter be set in order to notch out these tones? Would a band-limited comb filter be preferred for this application? Why?

5. Convert the following (using 8-bit positive binary numbers).
   a. 103 decimal to binary
   b. 17 decimal to binary
   c. 10010011 binary to decimal

6. A 6-bit A/D has a range of 0 to 8 volts. The sample-and-hold amplifier has an analog voltage present of 2.85 volts. What is the A/D binary output number?

11. ADAPTIVE NOISE CANCELLATION

Adaptive noise cancellation is illustrated in Figure 11-1. This process, commonly referred to as *two-channel* (2CH) *adaptive filtering* or *reference cancellation*, can be very effective in removing radio and television (TV) interference on a room microphone. In the figure, the talkers' voices are masked by the TV audio. This interference can subtracted from the microphone audio using a two-channel adaptive filter if the radio or TV audio is available as a reference.

A two-channel adaptive filter has two inputs, one for the mic audio and one for the noise reference audio, and a *single* output. It is ***not*** a stereo 1CH adaptive filter but operates on a completely different principle. A 2CH adaptive filter is capable of canceling *any type of noise*, even radio/TV voices masking the talkers.

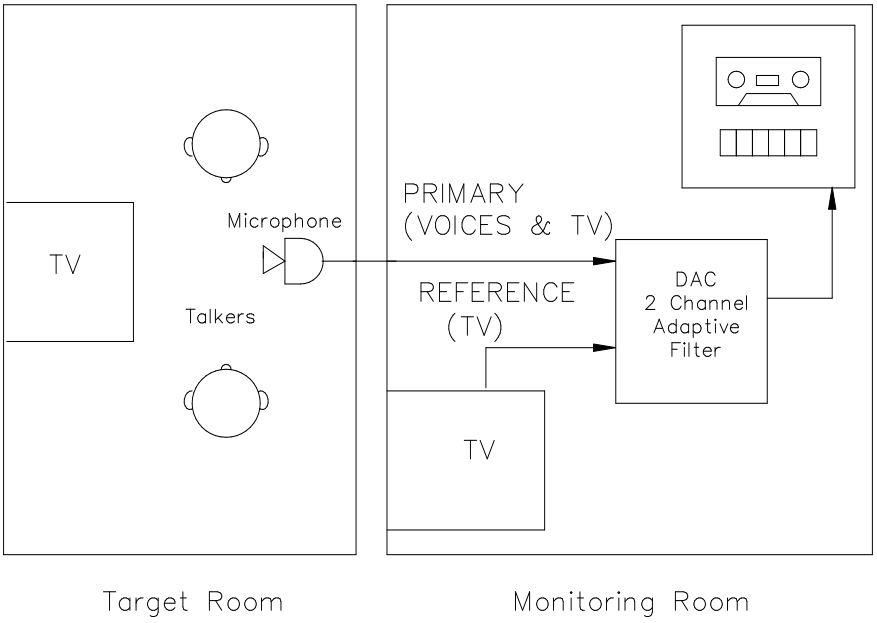


Figure 11-1: Two-channel Adaptive Filter Illustration

To understand a 2CH adaptive filter, consider Figure 11-2. Here a room microphone collects both voices and TV audio. Outside the room, audio is collected from a second, "reference" TV that is tuned to the same program. The reference TV audio is passed through an electronic simulation of the room and thereby made to precisely resemble the acoustic TV audio at the microphone. This modified audio, TV', is then subtracted from the microphone audio resulting in

$$\text{Voices} + \text{TV} - \text{TV}' = \text{Voices}$$

since TV' is identical to TV.

Note that only the TV audio is cancelled. The target audio is not input to the simulator and thus is unaffected by the process. Only the interfering TV audio is cancelled.

If the room had perfect sound treatment and produced no reflections, the room simulator would just need to delay and attenuate the reference TV audio to account for the acoustic path from the room TV to the microphone.

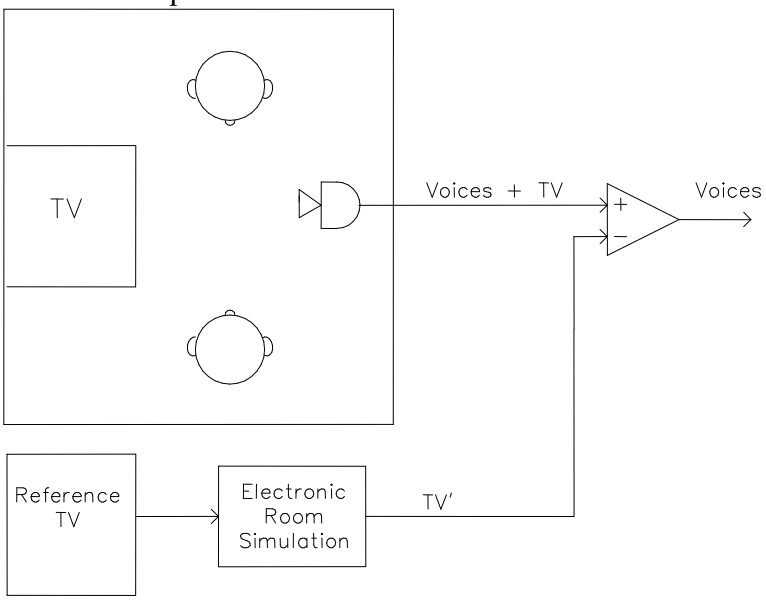


Figure 11-2: TV Cancellation Model

In real situations, however, the room's *transfer function* includes not only this direct path but also many reflected paths. Thus, the electronic room simulation becomes very complex. To manually design a simulator by measuring the room and surface reflection and absorption characteristics is virtually impossible. An automatic solution is therefore required.

The electronic room simulation is actually an FIR filter. The coefficients for this filter are derived using an adaptive filter algorithm similar to the 1CH adaptive filter discussed in Chapter 10.

To understand the 2CH adaptive filter, consider Figure 11-3. Microphone audio $V + N$ is delayed before reaching the differencer. The delay is noncritical and is usually set to a few milliseconds to give the TV reference audio a head start. The adaptive filter will automatically compensate for any excessive delay as long as it does not exceed the total length of the filter.

The TV audio is passed through the adaptive FIR filter, resulting in $\tilde{N}$, an estimate of the acoustic TV audio at the microphone. Since the voices V do not appear on the second input, an estimate of V cannot be produced by the linear filter.

The noise estimate $\tilde{N}$ is subtracted from the microphone audio $V + N$ to produce the error, or residue, signal E as shown by the following formula:

$$E = V + N - \tilde{N}. \qquad (5)$$

E is minimized in an LMS sense by the adaptation process in the same manner as in the adaptive predictive deconvolver of Section 10.3.2.2. The noise N is reduced without distorting the desired signal V.

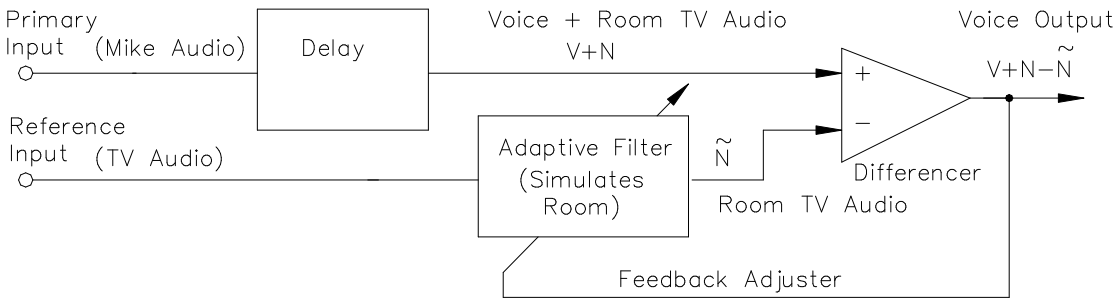


Figure 11-3: Adaptive Noise Cancellation

The residue signal is also fed back to the adaptation processor, which continually adjusts the filter coefficients as described by Equations (3) and (4) of Chapter 10.

The adaptive noise canceller cancels noise components common to the two audio inputs. Time correlation is not required; even random white noise may be canceled. The main requirement is that the microphone interference audio and reference audio be correlated with each other. This implies that they originate at the same source, nominally a point source. Distributed noise sources may be reduced in many cases by careful location of the reference and primary microphones.

As a rule, 2CH adaptive filters must be large filters to simulate all of the acoustic paths traversed by the TV audio. Consider Figure 11-4.

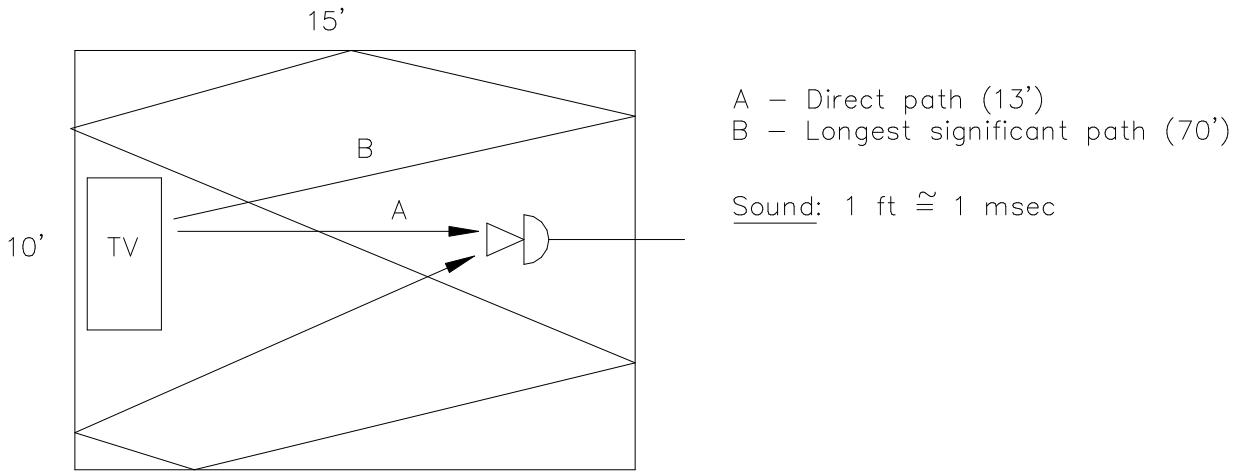


Figure 11-4: 2CH Adaptive Filter Size Considerations

The largest *significant* TV audio path is B, which is 70 feet long. Paths longer than this exist but have very low energy. In order for the adaptive FIR filter to simulate this path, it must have at least 70 msec of memory. Referring back to Figure 10-4, a filter with N taps has N delays, each separated in time by one sample interval.

If our filter has a sample rate of 12500 samples per second, each delay in the FIR filer holds

$$\frac{1}{12500} = 80\ \mu\text{sec}$$

of audio. The number of delays required, and hence the minimum filter size, is

$$\frac{70\ \text{msec}}{0.080\ \text{msec}} = 875\ \text{taps.}$$

Thus, an 875-tap filter is required. If the sample rate is higher, the filter size must also be larger.

A good rule of thumb to use when considering how large a filter is required to cancel all the significant TV audio paths is to take the largest dimension of the target room (in feet) and multiply by five. For example a 20' by 15' room would require a

$$20\ \text{ft.} \times 5\ \text{ms/ft.} = 100\ \text{ms}$$

filter to achieve good reduction of the TV audio. At a sampling rate of 12 kHz, this delay would correspond to a 1200-tap filter.

Two-Channel Adaptive Filtering
Dos and Don'ts

- **DO** use DATs for post-filtering.
- **DON'T** pre-filter mic (primary) or reference audio.
- **DON'T** use AGC on mic of reference audio.
- **DO** use a separate reference input for each independent noise source (stereo music excepted).
- **DON'T** allow reference audio level to fall below primary level.
- **DON'T** use excessive reference delay (best to use only 5 msec primary delay).
- **DON'T** adjust input levels unnecessarily (filter convergence affected).
- **DO** observe coefficient display where acoustic / transmission delays are uncertain.

EXERCISES

1. The longest reverberation path in a room (having significant energy) is 250 feet. What is the duration of an impulse in the room? What size two-channel adaptive filter sampling at 10 kHz would be required to cancel this path?

12. AUDIO ENHANCEMENT PROCEDURES

There is no single optimal methodology for enhancing audio tape recordings; different examiners may utilize different approaches and achieve essentially the same results. This section discusses both recommended methodology and the application of individual instruments.

12.1 Enhancement Methodology

Figure 12-1 is a flow chart that gives the sequence of enhancement steps that may be used in processing a noisy audio recording, whether tape-based or file-based. This chart is based on information given in Bruce Koenig's classic audio enhancement paper[4], which is included in Appendix D of this document. The chart has recently been updated to incorporate software-based processing tools in addition to hardware, owing to advances in technology since the paper was written.

But regardless of the technology that is used, forensic audio enhancement should always focus on the three critical voice qualities discussed in Chapter 4:

- intelligibility
- identification
- prosody

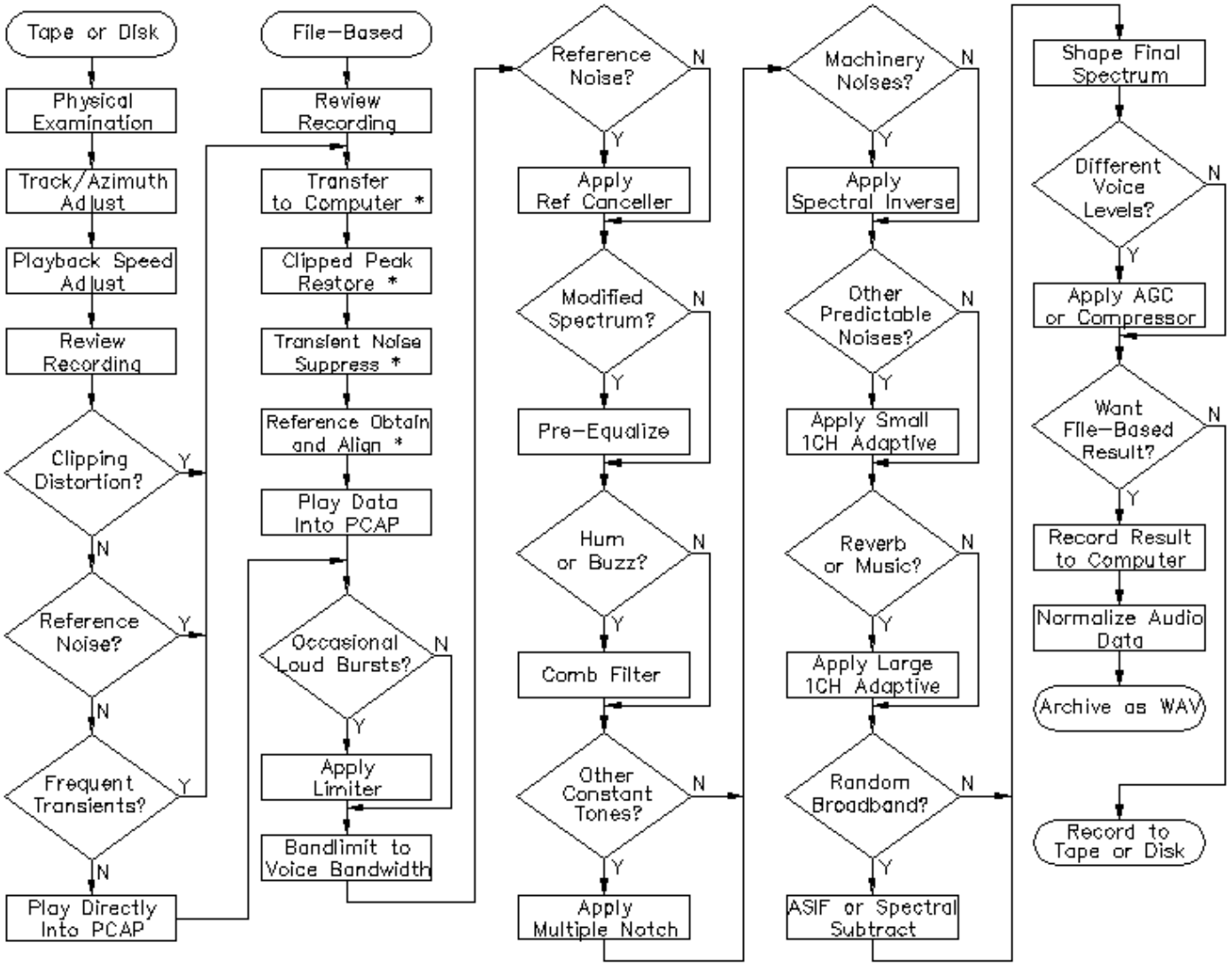
The goal of enhancement is to improve these qualities. A procedure which reduces any of these qualities by improving the *perceived quality* of the audio should be avoided. Some speech enhancement techniques do focus on making the audio clean, natural, and pleasant-sounding, but these are not the goals of forensic audio enhancement. Nonetheless, often the sound quality is also improved while enhancing the voice. This is especially useful when transcribing tape recordings, where disturbing artifacts increase transcriber fatigue and impair accuracy.

The generally accepted philosophy of tape enhancement includes the following five steps:

1. Optimize playback of recording.
2. Remove higher energy noise components.
3. Remove smaller energy noise components.
4. Reshape filtered voice spectrum.
5. Equalize voice levels.

Koenig further refines the process to 12 steps in his paper; again, due to changes in technology, DAC has recently built upon Koenig's process to include as many as 24 steps, depending upon the type of media being used and the types of noises and effects that are encountered.

[4] Bruce E. Koenig, "Enhancement of Forensic Audio Recordings," *J. Audio Eng. Soc.*, Volume XXXVI, No. 11.



* Require Additional Software-Based Tools

Figure 12-1: Enhancement Methodology Flowchart

The following comments refer to Figure 12-1. Some steps only apply to a tape or disk recording, while others apply to file-based media. In some cases, it may be necessary to temporarily convert audio from a tape or disk to a digital file for the application of software-based processes, the most useful of which are also described in these steps.

Physical Examination – The tape is inspected to determine whether it can be played back without damage. Obstacles to tape movement are corrected. Also, in some laboratories, the tape itself is subjected to ferrofluid analysis in order to (a) determine whether the tape is an original or a copy and (b) confirm the track configuration. Typically, a Freon-based compound known as "Magna-C" is used for this examination.

Track/Azimuth Adjust – The appropriate tape playback machine (track configuration and speed selection) is selected and azimuth (head alignment) is adjusted. On tape machines with adjustable azimuth, the screw or knob is adjusted until high-frequency response is maximized.

Playback Speed Adjust – The playback speed is adjusted to compensate for any speed errors introduced during recording, such as analog tape motor powered from a weak battery. An FFT spectrum analyzer is very useful in locating 50/60/400 Hz hum, dial tones, or any other *reference* signal. Speed should be checked and corrected across the entire recording.

Review Recording – Whether the recording is tape-, disk-, or file-based, the forensic technician should listen to the entire original recording prior to applying any enhancement filtering; this accomplishes two purposes, which are 1) to assess the noise problems that are present, and 2) to sufficiently familiarize the examiner with the recording so that the enhanced copy can be certified to be a true and accurate copy of the original.

Determine Whether Clipping Present – Encountering "broken speaker" sound in loud passages during the audio review would indicate that the recording device input level was set too high when the recording was made, resulting in "clipped" audio waveform peaks that might be treated with software tools for improved audio intelligibility[5].

Determine Whether Reference Noise Present – Recordings, for which voice intelligibility is degraded by identifiable, background audio material such as a music CD, a DVD movie, or a television or radio program, might be improved by Reference Noise Cancellation, *if* a digital copy of the interfering audio can be obtained. If the reference material is available, it can be aligned with the recording using software tools, and the two tracks can be played simultaneously through a Reference Noise Canceller for improved voice intelligibility[5].

Determine Whether Frequent Transients Present – "911", "999", "116", etc. tape recordings often have relatively loud computer keyboard "clicking" noise in the background, due to the

[5] For tape- or disk-based audio, it will first be necessary to first transfer the audio onto a computer hard disk via soundcard or other audio I/O device. For analog material, digitize the audio at 44.1kHz sample rate with 24-bit resolution for maximum signal integrity; for digital material, utilize the digital I/O, if available, and transfer the audio data at its original sample rate and data resolution for best results.

dispatcher, sometimes frantically, typing information as the telephone conversation is taking place. This directly degrades the intelligibility of the recording for forensic purposes, and also prevents adaptive audio filtering from getting the best results in reducing any far-end background noises that might be present. Software tools can be used to selectively compress the sections of audio in which the loud clicks are present, and then digitally amplify the remaining portions for improved intelligibility[5].

Clipped Peak Restore – If treatment of clipped peaks in the recording is desirable, it must be done as the very first enhancement step, prior to applying any other filtering. Clipped peak restoration software tools, such as those found in Sound Forge and Adobe Audition, generally operate on the mathematical principle of *spline curve estimation*; in this process, the desired segment of audio data is analyzed to determine the maximum sample value, and any short segments of audio that remain close to this maximum value will be replaced by new segments that more naturally follow the curve. Normally, for clipped peaks this means that waveform crests will be restored, which will always tend to increase the maximum sample value of the peak-restored audio. Therefore, a good tool will always allow for simultaneous digital attenuation of the peak-restored audio, so that sufficient headroom is provided to eliminate the previous overload condition.

Reference Noise Obtain and Align – Prior to utilizing Reference Noise Cancellation, digital reference material (e.g., an audio CD) must first be identified, obtained, and aligned with the buffered original audio using software tools. Only the reference material, not the buffered original audio, should be adjusted in any manner by the software during the alignment procedure. First, the reference material needs to be imported into the software at the same sample rate as the buffered original audio; in the case of an audio CD, it is recommended that the necessary tracks be directly "ripped" from the CD by the computer, to prevent losses due to analog conversion, etc. Next, corresponding marks need to be identified near both the beginning and the end of the two recordings, and the precise time difference down to 1-sample accuracy needs to be measured using the software. The time base of the reference audio then needs to be "stretched" to match that of the original audio, but without trying to preserve the "pitch" of the reference audio. Finally, the stretched reference needs to be time-aligned with the buffered original, such that corresponding marks in the reference occur approximately 10-20ms prior to those in the buffered original.

Transient Noise Suppress – If the audio is temporarily buffered on the computer, an additional step can be taken to reduce frequent loud transient sounds prior to noise-reduction processing. Software-based compressors can be selectively applied to strong transient noises in the audio, such as keyboard clicks, gunshots, etc., in order to suppress these noises and allow the level of the remaining audio to be amplified using digital gain.

Bandlimit – The audio signal above the voice's upper limit frequency (approximately 5 kHz) is attenuated to remove extraneous noises. Highpass filtering to remove the audio below 200 Hz is also recommended. This bandlimiting reduces low frequency rumble, tape hiss, and acoustic effects without impairing voice information.

Reference Noise Cancel – Assuming that the original audio is buffered on the computer, possibly with transient noise suppression supplied, and the reference material has been obtained and aligned with the buffered original audio, the Reference Canceller (or "2CH") adaptive filter can be used to automatically cancel the reference noise from the original audio. Care should be taken to ensure that the original audio is applied to the left (primary) channel of the PCAP, and that the aligned reference is applied to the right (reference) input; if these are accidentally swapped, and the routing to the PCAP is done using the digital I/O, it is possible to use the PCAP's Channel Swap feature to correct this error and make the cancellation function. Sufficiently large filter size should be utilized in order to maximize cancellation of reverberated reference noise; however, large values for adapt rate should be avoided to prevent filter "crashes" from occurring.

Pre-equalize Spectrum – Often the acoustic environment, microphone system, channel, or recording device will grossly modify the signal spectrum. This is especially true when concealed microphones are used, as is often the case in law enforcement recordings. Using a spectral equalizer or graphic filter, the spectrum can be reshaped to make it flatter (reducing dynamic range) and therefore more intelligible.

Hum and Buzz Reduction – AC power interference is very common to voice recordings. This noise consists of a sequence of harmonically-related tones. A comb filter is specifically designed to remove these harmonic tones, and should be used; however, a subsequent 1CH adaptive filter can help clean up any residual hum or buzz that might remain after comb filtering.

Constant Tone Reduction – Any tones present should be identified in frequency using a spectrum analyzer and then removed by a multiple notch filter. If the tones are moving or if tones are still present, a subsequent 1CH adaptive filter is recommended for best results.

Machinery Noise Reduction – Because they are readily identified by spectral analysis, constant machinery noises, such as air conditioning, fans, compressors, etc. can be effectively and precisely reduced by spectral inverse filtering. Using the built-in spectrum analyzer of the filter, a long-term spectral measurement should be captured and then used to build a precise equalization characteristic for reducing the noise. Digital gain incorporated into the filter can then be applied to the remaining audio to make it more intelligible.

Other Predictable Noise Reduction – The 1CH adaptive filter is a very effective tool for reducing any predictable noise components that might remain after spectral inverse filtering, most particularly when these noises are not constant but constantly changing.

Reverb and Music Reduction – Room acoustics, muffling, reverberations, and echoes can be reduced by deconvolutional filters including 1CH adaptive and spectral inverse (in the equalize-voice mode) filters. In the case of reverberation, a large filter size (generally greater than 1024 taps) will be required for the 1CH Adaptive filter to obtain the best results.

Random Noise Removal – If the audio is intelligible after all other required noise reduction has been applied, remaining low-level random noises may be suppressed by carefully applying either

a spectral subtraction filter, an automatic spectral inverse filter, or a multiband downward expander to the final audio. Care must be taken when adjusting the noise reduction settings to ensure that excessive values are not used; if they are, the intelligibility may be degraded and objectionable audible artifacts may be introduced.

Output Spectral Shaping – A graphic or parametric equalizer may be used to reshape the voice's final output spectrum, improving overall quality. This is a cosmetic step applied once all noise removal is completed.

Voice Level Equalization – When noise-reduced voices present have significantly different voice levels, an automatic gain control (AGC) or compressor-expander may be applied to reduce speaker volume differences.

Record Result to Computer – Particularly in cases where the ultimate audio product will need to be burned to CD or archived to some form of data file media, recording the enhanced audio back to the computer in WAV or other popular format may be required. This allows a "second crack" with the software tools in order to achieve the best final result.

Normalize Audio Data – In those cases where the filtered audio is re-recorded to the computer hard drive, it is often useful to *normalize* the audio data using software tools so that the loudest peaks are at a good level. Also, if the audio is not sampled at 44.1kHz, it will be incompatible with burning to audio CD; therefore, the software can be used for sample rate conversion, if needed.

12.2 Application of Enhancement Instruments

Implementation of an enhancement setup requires careful selection of the sequence of filters, limiters, equalizers, etc. to carry out the required task and the adjustment of these enhancement instruments.

As described in Section 12.1, the same type of instrument may be used to attack different types of noises, *e.g.*, a 1CH adaptive filter may be used as both as tone filter and a reverberation deconvolver. The same type of filter may be used at different stages in the enhancement sequence, *e.g.*, an equalizer may be applied to reducing banded noise in an earlier stage and to reshape the overall output spectrum as a final stage.

This section discusses the application and adjustment of various enhancement instruments as they might be applied to any stage in the enhancement setup. Table 11 summarizes these filters and processors, and the remainder of this section details them.

Table 11: Forensic Audio Filtering

Filter / Processor	Application
Clipped-Peak Restoration	Non-linear “clipping” distortion (“broken speaker” sound)
Transient Noise Compression	Keyboard clicks, gunshots, or any other repetitive, “pulsing” noise
Limiter	Overload protection
Reference Canceller	Background Noises for which Digital Reference is Available (e.g. music CD)
Comb	AC Mains Hum, Harmonic Tones
Multiple Notch	Tones
Graphic	Multipurpose, precision reshaping
Adaptive Deconvolution	Reverberation, Predictable Noises
Spectral Inverse	Reverberation, spectral “coloration”
Equalizers (Parametric and Graphic)	Spectral reshaping
Spectral subtraction	Limited broad-band random noise
Compressor/AGC	Voice Level Equalization

12.2.1 Tape Speed Correction

Tapes are often recorded at an incorrect speed. Errors up to ±10 percent are not unusual with field machines that have been in extensive service. The playback speed during copying and enhancement should be adjusted to match the original record speed.

An FFT spectrum analyzer can be very effective in identifying the actual record speed using reference frequencies that might exist in the audio. These may include

- Power (50, 60, and 400 Hz),
- Dual Tone Multiple Frequency (DTMF) (tone dialing),
- Busy and dial tones, and
- Other special known frequency tones.

Touch-tone telephones use two simultaneous tones to represent each key on the touch pad with a system called Dual Tone Multiple Frequency (DTMF). When a key is pressed, the tone of the column and the tone of the row are generated. The frequencies were chosen to avoid harmonics. All telephone sets use the digits 1-0 and the * and # keys, while U.S. military and telephone company phones also use the A, B, C and D keys. The frequencies for the DTMF tone pairs are given in Table 12.

Table 12: DTMF Tone Pairs

Signal	Low-Frequency Component (Hz)	High-Frequency Component (Hz)
1	697	1209
2	697	1336
3	697	1477
4	770	1209
5	770	1336
6	770	1477
7	852	1209
8	852	1336
9	852	1477
0	941	1336
*	941	1209
#	941	1477
A	697	1633
B	770	1633
C	852	1633
D	941	1633

Example: A cassette recording is found to have steady spectral lines at 224, 280, 336, and 392 Hz. The tape is known to originate in North America. What was its record speed?

Solution: The lines are harmonically related with a fundamental frequency of 56 Hz. The first three components are missing because of the poor low frequency response of the recorder. The playback speed is adjusted such that the *highest distinct harmonic* matches its nominal frequency. In the example, 392 Hz is the seventh harmonic of 56 Hz. That spectral line is set to 7 × 60 Hz = 420 Hz by increasing the playback speed.

12.2.2 Clipped Peak Restoration

Whenever loud audio in the recording produces "broken speaker" sound, even though the listening equipment is set to a proper volume, this is an indication that the inputs to the recording equipment were likely overdriven when the recording was made. Transferring the audio to a computer and observing the digitized data with an editor display will confirm that clipping has occurred, and then a software-based clipped peak restoration tool can be used to correct the clipped peaks *without* affecting the surrounding audio data that is not clipped. This should be done *as the very first step in the enhancement, before any additional noise filtering is attempted,* because filtering will make the clipped peaks less distinguishable to the software tool. Properly done, the effect will be to reduce the harmonic distortion that produces the broken speaker sound,

and will allow subsequent noise reduction filters to do a better job of improving the remaining audio.

12.2.3 Transient Noise Suppression

Transient noises (those that are not constant, and are very short in duration) can affect the intelligibility of the recording in a similar manner to near/far party effect; the loud sounds overwhelm the ears, and the lower-level audio of interest cannot be distinguished. Transferring the audio to a computer, then first applying clipped peak restoration if necessary, will allow a software-based audio compressor tool to be selectively applied to the portions of the audio where the loud transients are present, in order to reduce their volume without affecting the volume of the remaining audio. Finally, digital gain (or normalization) can be applied to the transient-reduced audio in order to make the remaining audio more distinguishable, both to the ears and to any subsequent noise reduction stages that might further improve intelligibility.

12.2.4 Input Limiting

An input limiter should be used wherever an *occasional* loud input burst, capable of overloading the subsequent processing devices, is possible.

The attack time, *i.e.*, the interval required to respond to a loud sound, should be short, on the order of 1 msec. The release time, *i.e.*, the interval required for the limiter to return to normal gain following a loud sound, should be on the order to 250 to 1000 msec. The shorter interval enables rapid recovery following a loud sound, but it may add a modest level of distortion during recovery.

12.2.5 Upper and Lower Bandlimit Filtering: Lowpass and Highpass Filters

Sounds generated by human speech occur in the frequency spectrum from 100 hertz to above 10,000 hertz. The entire voice spectrum is not required for speech intelligibility and speaker identification. The voice frequency spectrum from 300 to 3,000 hertz is sufficient to reliably transmit voice intelligibility over a telephone line. This is the frequency spectrum that is transmitted by the telephone company, most narrow band RF transmitters, and microcassette recorders. If this reduced voice frequency spectrum (300–3,000 Hz) is degraded by noise, intelligibility problems usually occur. A person listening to a live conversation has the full frequency spectrum available to aid in speech intelligibility and discrimination against interfering noises. When a telephone system, a small RF transmitter, or a miniature cassette recorder is used to capture a conversation, much of the voice frequency spectrum is lost. Therefore, a covert recording of a conversation is oftentimes less intelligible than the live conversation.

As a general rule, sounds above and below the voice frequency spectrum should be removed by lowpass and highpass filters, respectively, to enhance speech intelligibility. A bandpass filter incorporates both a lowpass and highpass filter for simultaneously removing interfering noise from the voice information above and below the restricted voice frequency spectrum. A bandstop filter removes banded noise within the voice spectrum and also incorporates a lowpass and highpass filter. A notch filter is a very narrow bandstop filter that is normally used to remove discrete tones.

The experienced examiner strives to retain all frequencies associated with the human voice. The rule is not to remove any signal components which reduce voice information even if doing so results in better sounding audio.

12.2.5.1 Lowpass Filtering

The upper bandlimit of speech on a recording is often known from the specifications of the device or circuit. Such information is generally known about telephones, body transmitters and recorders, police and other communications systems, microcassette recorders, and concealment devices.

If unknown, the upper limit of voice frequencies can be determined using an FFT spectral analyzer. High frequency fluctuations in the spectrum synchronous with audible speech may be used to establish upper voice frequency limits.

The upper band limiting may be addressed with a lowpass filter adjusted to the voice's upper limit frequency. Many DAC filters have selectable bandwidths, eliminating the need for lowpass filtering. Selecting the filter processing bandwidth to match that of the voice audio has the additional advantage: all of the processor's resources are focused on the actual voice bandwidth.

The lowpass filter frequency adjustment setting for optimal voice intelligibility may be determined by starting at 3,000 Hz and progressively increasing the upper frequency limit above 3,000 Hz through several steps using a step size of 100 to 250 Hz. At each step, critically listen to the speaker's voice to determine whether speech intelligibility has been improved.

A common problem the tape analyst encounters is making the trade off between better sounding audio and better intelligibility. When an interfering noise is reduced, the sound is usually improved; however, the conversation must be aurally reviewed to verify that the intelligibility of the talkers was improved (*i.e.*, additional words can be understood). Again, the primary principal is never to remove a noise that reduces speech intelligibility.

12.2.5.2 Highpass Filtering

Energy below the voice's lower cutoff frequency usually consists of room rumble and low frequency noises. A 200 Hz highpass filter is usually adequate to attenuate most of this noise. Since virtually no voice information is lost in removing the lower 200 Hz, *a highpass filter may, therefore, be used for virtually all enhancement processing.*

A more precise method of adjusting the highpass filter's cutoff frequency is to start at 300 Hz and progressively decrease the lower frequency limit below 300 Hz through several steps using a step size of 10 to 50 Hz. Stop at each step to critically listen to determine whether speech intelligibility has been improved. *Highpass filtering speech at cutoff frequencies greater than 300 Hz is not recommended.*

12.2.5.3 Bandpass Filtering

A bandpass filter is a serial combination of a lowpass and a highpass filter.

A significant advantage of DAC's filters is their ability to control the rate of rolloff at the cutoff frequencies. Very steep rolloffs, *i.e.*, large Qs, can easily be achieved, but these should be used sparingly to avoid adverse effects to the voice (*e.g.*, ringing and unnatural sounds). Again, critical listening is crucial to producing the best product. A good procedure is to change only one setting at a time and to stop and listen to determine the effect of this change. Continue changing this control until the voice intelligibility (not listenability) ceases to improve.

In general, the tape enhancement will start at an initial lower and upper setting of 300 and 3,000 Hz. The operator will use a step size of 100 to 250 Hz at the high end and a step size of 10 to 50 Hz at the low end and adjust these settings to produce maximum speech intelligibility. The analysis is always critically reviewed using single step changes. If two changes are made concurrently, the improvement or degradation of a single step change is difficult to ascertain.

12.2.6 Within-Band Filtering: Bandstop, Notch, and Comb Filters

Bandstop, notch, and comb filters attack noise within the voice bandlimits. Spectrum equalizers, 1CH adaptive filters, and spectral inverse filters are also used and will be subsequently discussed.

Whenever the examiner is operating within the voice frequency bandwidth, extreme care must be taken to remove the minimum amount of in-band energy. The bandwidth and depth of each spectral slice should be manually adjusted. After each adjustment, the examiner should aurally review the results and select only filtering operations that not only reduce the noise but actually increase talker intelligibility.

12.2.6.1 Bandstop Filtering

A bandstop filter consists of a lowpass and a highpass filter summed together to form the output audio. For example, a highpass filter passing audio above 2500 Hz (its F_C) and a lowpass filter passing audio below 2000 Hz would have a stop band from 2000 to 2500 Hz. The cutoff frequency for the highpass must be higher than the cutoff frequency for the lowpass; otherwise, they would overlap and no stopband would exist.

Like bandpass filters, DAC's bandstop filters have adjustable cutoff frequencies, rolloff, and stopband attenuation. Normally, the FFT spectrum analyzer is used to identify the section of the spectrum, and the bandstop upper and lower cutoff frequencies are adjusted to reduce the noise level. Like adjusting other filters, the operator should be sensitive to producing maximum intelligibility. Severe amputation of the audio spectrum may greatly reduce the noise but may also inadvertently impair the voice.

12.2.6.2 Notch Filtering

Notch filters are narrow bandstop filters. Their stopbands in analog filters are as small as 50 Hz and in digital filters are as small as 1 Hz. A parametric equalizer may also be adjusted to act as a notch filter. The notch frequency and width are normally adjusted to remove a tone from the audio. DAC filters also have a notch depth adjustment. An FFT analyzer is most useful, but an experienced operator can often adjust the notch frequency by ear. If multiple tones are present, requiring adjusting multiple notch frequencies, a spectrum analyzer becomes more necessary.

The notch width, measured in Hz, is adjusted to span the frequency bandwidth occupied by the tone. If a digital recording or live signal is being filtered, the tone is normally extremely narrow in bandwidth. If the audio is from an analog recorder, wow and flutter will cause the tone to jitter over a small frequency range; the notch bandwidth must be increased in this case.

The notch depth is useful in controlling the amount of spectrum being removed. Notching the tone down just below the voice level is often quite satisfactory and has minimal detrimental effect on the voice signal.

12.2.6.3 Comb Filtering

The comb filter is effective in reducing harmonic interference hum as might be introduced by power line coupling. Operating at 60 Hz, the comb filter places a null at 60 Hz and all of its harmonics (120, 180, 240, … Hz), thus canceling powerline-generated noises. Under certain circumstances, such as motor brush arcing or fluorescent light noise, strong 60 Hz harmonic components along with random noise spikes are produced. These random, *hash* components in these situations may not be completely removed.

A comb filter often provides the following features:

- Fundamental notch frequency control,
- Upper notch limit control,
- Notch depth adjustment, and
- Odd, even, and all harmonic selection.

The fundamental notch frequency control permits placing the fundamental notch frequency over a broad range of frequencies, typically between 40 and 500 Hz. This comb filter is usually recommended where the hum component frequencies are stable. If the precise fundamental frequency is unknown, the frequency is adjusted until a null is achieved in the hum. If the fundamental hum frequency is varying, as is often the case with speed fluctuations in poor tape recordings, then the comb filter should be followed by a 1CH adaptive filter.

The upper notch limit frequency specifies the frequency of the highest complete notch. Above this frequency, the notches rapidly diminish in depth. Typically, at 400 Hz above the upper notch limit frequency, the filter transfer function is essentially flat. This notch limit feature allows the operator to place notches over only the range where the hum harmonics exist. Comb filters characteristically introduce mild reverberations. By limiting the notches to the range of significant hum energy, usually less than 1000 Hz, the reverberation artifact can be reduced. The recommended adjustment procedure is to decrease the notch limit frequency from 3000 Hz until just before the audible hum components begin to increase. Seldom will a full 3000 Hz limit be required.

The third comb filter control is the notch depth adjustment. This value may be adjusted over a range from 10 to 50 dB. Excessive notch depth may contribute to degradation in signal quality. The recommended procedure is to start with the maximum depth, 50 dB, and to reduce this value until hum components begin to increase.

The final comb filter adjustment is the selection of odd, even, or all harmonic notches. Due to waveform symmetry, the hum may have only odd harmonics (60, 180, 300, 420 Hz, etc.) or even harmonics (120, 240, 360, 480 Hz, etc.). Generally, all harmonics are present. Using an FFT spectrum analyzer or merely selecting Odd, Even, and All with careful listening will establish the best choice. As with all within-band filtering, introduce no more notches in the voice spectrum than is absolutely necessary.

A comb filter is a reverberation-introducing device. A 1CH adaptive filter often follows a comb filter to both reduce the reverberation introduced and to clean up any remaining hum component.

12.2.7 1CH Adaptive Filter

1CH adaptive filtering, also known as adaptive predictive deconvolution, is used for removing time-correlated noise components from voice signals. Problems attacked include noises such as

music, tones, and hum, and acoustic effects such as echoes and reverberations. As the filter is adaptive, it continuously tracks changing noise characteristics and reduces the noise energy. Music, engine sounds, and other dynamic noises can also be significantly reduced by a 1CH adaptive filter.

The functional block diagram of a 1CH adaptive filter is given in Figure 12-2. The input signal is processed by a large adaptive predictive digital filter. This filter predicts the signal slightly in the future (based on the prediction span) using the recent history of the signal stored in the filter. The *predicted* signal is then subtracted from the actual signal, resulting in the *residue*. Statistically random components are not predictable, but time-correlated components are. Voice signals are principally random and appear in the residue output. Noise components are often time-correlated and are therefore accurately predicted; thus, they are cancelled very effectively in the subtraction process.

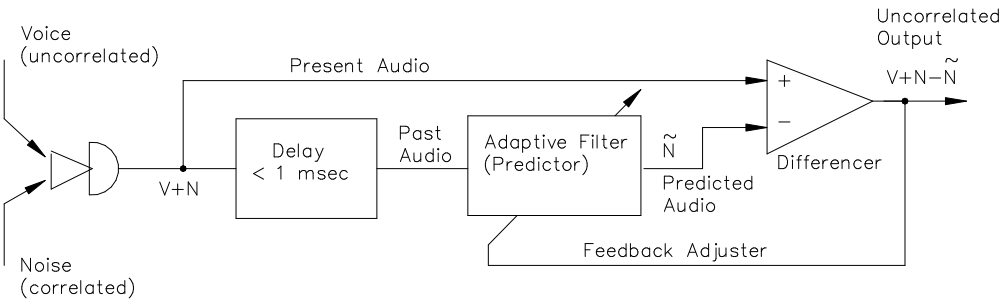


Figure 12-2: 1CH Adaptive Predictive Deconvolution

Adaptive filter controls include filter size, prediction span, and adapt rate. The filter size determines the amount of recent history that is stored in the filter. Longer filters generally predict, and hence cancel, noises better. This is especially true for echoes and reverberations; however, with simple noises such as tones, a filter size that is too large may introduce a hollow sound in the output audio.

Normally, the filter size is initially set to its maximum value. The filter size is then reduced one step at a time, and the filter is allowed to converge for each step. Each step can be a 10 to 25 percent size reduction. Once the noise level starts to increase, the next larger filter size is selected. As a rule, this adjustment is not critical; however, one may wish to use the smallest acceptable filter size to reduce music and other dynamic noises. Smaller filters will accept larger adapt rates, and hence will track more rapidly, without crashing (filter divergence resulting in a noisy output).

The prediction span is the time, in milliseconds (msec), into the future that the signal is predicted. Longer prediction spans result in more natural sounding voices, whereas shorter predictions enable the filter to attack less correlated noises. In effect, this control enables the operator to compromise between voice quality and noise cancellation. If only stable noises, such as hum or acoustic effects, are present, then a longer prediction span may be used without penalty.

The adapt rate control specifies the speed at which the adaptive filter adjusts itself. The control is actually adjusting the μ adapt coefficient. Slow rates are suitable for stationary noises such as hum and reverberations. Faster rates are better able to track dynamic noises such as popular music. On the other hand, faster rates may also attack correlated components of the voice, such as sustained vowels. Therefore, the adapt rate should be adjusted to the slowest rate that adequately tracks the noise and minimizes the adverse effects on the voice.

Too fast an adapt rate will cause the filter to diverge or crash (output audio becomes intense noise). Should this occur, clear the filter and reduce either the adapt rate or the filter size. Smaller filters can tolerate faster adapt rates without crashing. The adaptive filter's Reset or Clear push button will reinitialize the filter following a crash.

The adapt rate may be fixed or normalized. The stability of an adaptive filter is affected by the combination of the adapt rate coefficient μ, the filter size, and the input audio level. If a constant (*i.e.*, fixed) μ is used, it must be small enough for the filter to remain stable (*i.e.*, not crash) at the loudest input level. Normalization automatically reduces μ for louder input levels and increases μ for softer levels. Effectively, normalization *matches* the μ to the audio level and allows for faster filter adjustment. The normalized (or "Auto") adapt rate is usually recommended.

The user can control when the filter adapts with *conditional adaptation*. The adapt control may be set to adapt continuously (Always) or not to adapt (Freeze). When the adapt process is frozen, the filter continues to reduce noise on the signal, but it will not readjust itself to new or changing noises. Conditional adaptation modes permit the user to enable adaptation when the input or output signals exceed or fall below an adjustable threshold level. If audio is interrupted occasionally by a loud noise such as a door slam, setting the adaptive filter to adapt when the audio is below a threshold will prevent the filter from readjusting itself to the door noise. In this case, the threshold would be set above the voices but below the loud door slam level.

Alternatively, if voices of interest are only sporadic, the filter may readjust itself to low-level background sounds in the voice gaps. To keep the filter solution intact during these gaps, the filter can be set to adapt during high-level (voices present) segments. The adapt threshold would be set to a level below the voices but above the residual background noises.

Because of its versatile nature in attacking many noises common to forensic recordings, a 1CH adaptive filter is often used by itself to enhance recordings. These simpler devices generally include a bypass switch for before/after comparisons and a 200 Hz highpass filter for reducing low frequency room noises and rumble on the output audio. Spectrum equalization is often used in conjunction with a 1CH adaptive filter to both *pre-equalize* and *post-equalize* the audio. This process is illustrated in Figure 12-3.

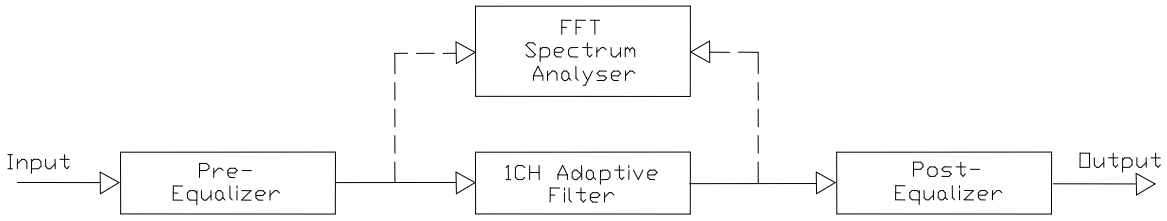


Figure 12-3: Adaptive Filter with Equalization

The purpose of pre-equalization is to spectrally flatten the input audio and to bandpass filter the signal minimizing out of band noise. If tones are present, the equalizer may also be used to notch them. An FFT spectral analyzer is used to adjust the spectrum equalizer. Pre-equalization is a coarse spectral adjustment and conditions the audio for the 1CH adaptive filter.

The 1CH adaptive filter is the heart of the enhancement sequence. It carries out automatic spectral adjustment and noise reduction. The 1CH adaptive filter also flattens the spectrum.

The post-processing spectrum equalizer reshapes the adaptive filter's audio to a more natural-sounding signal. Normally, high and low frequencies are rolled off to more resemble the voice spectrum. Post-processing is principally a cosmetic step. Care should be taken to not sacrifice intelligibility while making the voice sound more pleasing. An FFT analyzer is often useful in making these spectral adjustments, but generally the best results are obtained by adjusting by ear.

12.2.8 Spectral Subtraction Filtering

Spectral subtraction filtering is a popular method of "denoising" speech recordings; much like the proverbial "sledge hammer", they tend to go after all types of non-speech audio, regardless of whether they are predictable or random in nature.

Generally, they are *not* capable of extracting voices that are buried below the noise floor, because the noise floor is, in fact, what is removed by the filter, and anything below it will be removed or at least severely degraded.

Also, because most implementations of the algorithm are accomplished in the frequency domain, to reduce computational complexity, audible "birdy noise" artifacts will generally be produced by the filter, particularly when excessive noise reduction amounts are applied.

Therefore, the general recommendation is to use the spectral subtraction filter as a "polishing" stage, after all other more selective noise reduction options have been applied. The minimal amount of noise reduction that provides satisfactory results should always be applied, in order that the output has maximum speech intelligibility and minimal audible artifact from the filtering.

12.3 2CH Radio/TV Removal

12.3.1 Application of a 2CH Adaptive Filter

Two-channel (2CH) adaptive filtering, also known as adaptive noise cancellation, can be very effective at removing radio and television audio from room microphones. This process is called noise cancellation because, unlike 1CH adaptive *filtering*, the noise is phased and cancelled. Talkers often move close to a television (TV) or radio and increase its volume in order to mask their private conversations. Two-channel adaptive filtering can cancel this masking radio or TV sound, thereby rendering the desired conversation understandable.

The Figure 12-4 shows a target room, with talkers and masking TV audio, and a nearby room acting as a monitoring room. The target room audio consists of voices (the desired audio) along with undesired TV audio (the noise).

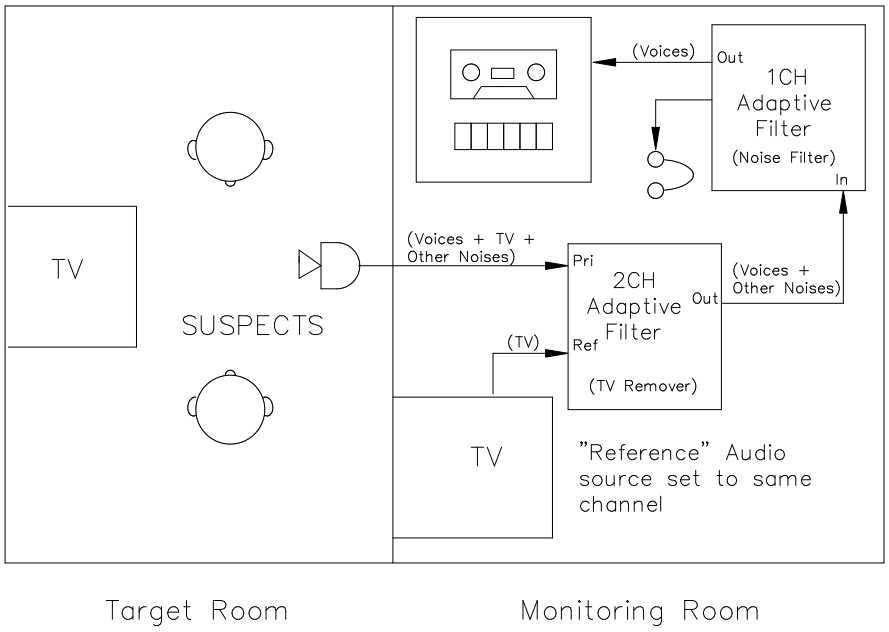


Figure 12-4: Adaptive Noise Cancellation

The target room's microphone signal is input to the adaptive filter's *primary* audio input, and the TV audio is connected into its noise *reference* input. Notice that the TV audio is derived from a second TV receiver in the listening post, which obviously must be tuned to the same TV channel. The two-channel adaptive filter develops an electronic model of the room and *filters* the reference audio, making it resemble the TV audio appearing at the microphone. This electronic model replicates the multiple acoustic paths by which the TV sounds reach the microphone in the target room. This *electronically filtered* TV audio is then subtracted from the microphone signal, canceling the TV audio components and leaving the voices intact. Unlike 1CH adaptive filtering, the microphone signal is not filtered; hence, the voice is not modified.

The two-channel adaptive filter carries out its adjustment automatically. It continuously fine tunes itself as acoustic paths change, such as when one of the talkers moves. The adaptive filter is extremely easy to operate once the procedure is understood.

The following nine points are suggestions on applying adaptive filters:

1. Filter Size – Large digital filters are normally used in 2CH adaptive filtering. Most rooms require at least a 1000-tap[6] order filter, but large areas and hard, reflective walls may require filters of 4000 or more taps.

The adaptive filter precisely models the acoustic paths in the room that the noise traverses from its source to the microphone. The paths include both the *direct* and *reflected* paths. The reflected

[6] The size of a digital filter is measured in both *taps* and *filter order*. These terms are used interchangeably.

paths are often numerous and quite long. The filter's *memory* is determined by the following equation:

$$\text{memory} = \frac{\text{filter size}}{\text{sample rate}}$$

As an example, a 7 kHz bandwidth filter having a sample rate of 16.5 kHz and 4096 taps would have

$$\frac{4096}{16500\ \text{Hz}} = 248\ \text{msec}$$

of memory. Since sound travels approximately 1.1 foot per msec (3 msec per meter), this filter can model up to 272 feet of reflected acoustic path.

Unlike 1CH adaptive filters, the maximum available filter size is normally used. If the filter is too small, reflected sound paths longer than the filter's memory would not be cancelled and would remain in the output audio.

As a practical consideration, 2CH adaptive filters are capable of 20 to 30 dB of TV audio reduction in a typical room. Acoustic paths that are lower in level than this generally are not removed due to arithmetic precision or subtle variations in path characteristics from air movement, temperature changes, etc. They are also probably masked by other room background noises. By measuring the *impulse response* of the room and determining the duration for the response to die out 30 to 40 dB, the minimum filter size can be specified.

Example:

A room's impulse response dies to a level 30 dB below its initial level in 185 msec. What is the minimum 2CH adaptive filter size required for filters with 5 kHz and 7 kHz bandwidth? Assume the sample rate is 2.5× the bandwidth.

at 5 kHz BW

sample rate = 5,000 Hz × 2.5 = 12,500 samples per second

filter size = 0.185 × 12,500 = 2,313 taps

at 7 kHz BW

sample rate = 7,000 Hz × 2.5 = 17,500 samples per second

filter size = 0.185 × 17,500 = 3,238 taps

Note that filter size requirements increase with bandwidth.

2. Adapt Rate – The tendency to increase the adapt rate to too large a value is common in this mode. In doing so, the filter adjustment corrections become rather large, which causes the output signal to modulate the coefficients. The resulting audio begins to sound hollow or reverberant. In the extreme case, too large of an adapt rate will cause the filter to crash resulting in output audio of random noise. Be aware of this effect of faster rate settings and decrease this setting if necessary.

When a 2CH adaptive filter is first switched on, a larger adapt rate may be used to expedite convergence. Once an acceptable level of cancellation has been achieved, the adapt rate may be reduced to a maintenance level.

3. Conditional Adaptation – Certain adaptive filters incorporate conditional adaptation; this feature allows the input or output signal level to specify when the filter is adapting and when it freezes. When frozen, the filter continues to cancel the TV audio but will not readjust itself to changing room conditions.

To use this feature the filter is initially allowed to converge to near maximum TV cancellation; this can take 60 seconds or more. Once cancellation is achieved, the adapt control is set to switch on adaptation when the output audio is below a threshold. This position is usually labeled "Output < Threshold" or "Residue < Threshold." The threshold is adjusted to freeze the filter when the talkers speak but to allow the filter to adapt during their silences. Since the residual TV audio is below the threshold, the filter will possibly achieve a better degree of cancellation. Final adaptation occurs when only the TV is present; the voices will not "confuse" the process.

4. Delay Adjustments – A 2CH adaptive filter usually has a delay line which may be placed in either the primary or reference input path. The purpose of this delay is to make sure that the reference TV audio reaches the differencer ahead of the corresponding audio from the microphone. See Figure 11-2. In Figure 12-4 the TV audio reaches the filter (reference) input immediately and reaches the primary input via a delayed acoustic path. No delay is required for this setup since the acoustic delay guarantees that the necessary conditions are met.

For most conventional applications, the delay line is placed in the primary input path and adjusted to 3 to 5 msec. By doing so, the delay requirements are always met, even if the microphone in Figure 11-2 is placed next to (or possibly behind) the TV's loudspeaker.

Additional discussion of filter delay is given in Section 12.3.2.

5. Input Level Adjustment – In order to avoid exceeding the adaptive filter's coefficient dynamic range, the reference audio level is normally adjusted to be at least as large as the primary audio level. The filter's bargraphs assist in meeting this requirement.

6. Post-processing – Two-channel adaptive filtering removes only the designated TV audio. Other noises and acoustic effects on the microphone signal remain. By following the two-channel filter with a one-channel adaptive filter, as shown in Figure 12-4, the remaining audio is

often enhanced, reducing undesired noises and acoustic effects. DAC offers several 2CH–1CH instruments.

7. Remote Processing – It is sometimes necessary to use a 2CH adaptive filter at a remote location. The microphone audio and sometimes the reference audio are communicated to the monitoring location via telephone, RF, or some special link.

As long as both filter inputs are received with minimal distortion, noise cancellation should be unaffected by this process. Unfortunately, certain communication channels introduce nonlinear distortion, severe bandlimiting, and AGC effects which do impair processing.

Nonlinear distortion introduces *new* frequency components that are not present in the original audio. These components are not cancelled. Band limiting of the microphone audio does not impair cancellation but does affect voice quality.

Automatic gain control (AGC) changes the microphone gain dynamically. Since no matching gain changes occur on the TV reference audio, the filter must compensate. Normally, the filter can not adapt as fast as the gain changes, and TV cancellation is thus impaired. Therefore, AGC should be avoided in this application.

8. 2CH Adaptive Filter with Previously Recorded Audio – In certain situations it may be necessary to simultaneously record the microphone and reference TV audio on to a stereo tape recorder and subsequently cancel the TV audio by inputting the two audio tracks into a 2CH adaptive filter. This process can be very successful but has two potential drawbacks.

First, by not testing the microphone and reference TV setup for 2CH canceling ability, bad microphone position or technical problems may not be detectable until after the audio is already recorded.

The second caveat involves the stereo recorder use. If a DAT is employed, adaptive noise cancellation results should be as good as working with live audio. The DAT does not suffer from wow and flutter, amplitude variations, crosstalk, or distortion. If an analog stereo recorder is used, these factors severely limit the filter's ability to cancel. Instead of potentially 20 to 30 dB of TV audio cancellation, processing analog recorded audio typically results in only 5 to 15 dB of cancellation. Open reel recorders at 7.5 or 15 ips usually perform better than slower speed cassette formats due to superior speed regulation and crosstalk specifications. The stereo audio record capability of a Hi-Fi VHS recorder is even better, achieving near-DAT quality with an effective tape speed of over 100 ips.

Carrying out 2CH adaptive filtering using two independent DAT recordings, one for the microphone and one for the TV reference audio, is more difficult but can sometimes be successful. Sound editing software, such as Adobe Audition, Sound Forge, or WaveLab, can often be used to time-align these independent recordings and play them back simultaneously through a 2CH adaptive filter for decent results. Using two independent analog recordings is

virtually impossible due to independent speed fluctuations (wow and flutter) of the two recorders.

9. Limiting Amount of Cancellation – In certain situations, it is appropriate not to cancel the TV audio completely. It may be appropriate to do a less-than-best job in order for the resulting audio to still have some background TV audio present. Otherwise, the result may be so good that doctoring or editing may be suspected.

Limiting cancellation is accomplished by allowing the filter to reduce the TV audio 10 to 20 dB and then *freezing* the filter. If the environment changes and cancellation drops, adaptation can be temporarily switched back on to return to the desired level of cancellation.

12.3.2 Coefficient Display and Interpretation

Most DAC 2CH adaptive filter products are fitted with a coefficient display feature for monitoring the internal operation of the digital filter. This feature continually displays the filter's impulse response, which is simply a graph of the filter coefficients and is very helpful for adjusting filter size and delay. It gives the operator valuable insight into the overall operation of the filter. This feature is very helpful in adjusting input delay in the multi-channel mode of operation and giving the operator insight into the overall operation of the filter.

DAC's personal-computer-based products (PCAP, MCAP, and MicroDAC IV) have the ability to display the coefficients on the computer's monitor.

12.3.2.1 Interpretation of the Impulse Response for a 1CH Adaptive Filter

A simple demonstration of the filter impulse response may be made by using the adaptive filter in the 1CH (adaptive predictive deconvolution) mode to cancel a sine wave. Apply a 100 Hz sine wave and permit the filter to slowly adapt, canceling that waveform. Observe the coefficient display and notice that it also is a sine wave. As the filter size is increased from its minimum size, the impulse response becomes longer, and the values become smaller. Also notice that if adapt rate is made larger, the displayed sine wave impulse response appears to decay exponentially; this is normal.

The Fourier transform of the impulse response is actually the spectral transfer function of the digital filter. A real-time spectrum analyzer may be used to display this function. The principal use of the impulse response in the 1CH mode, however, is in training the operator's intuition. Strong peaks in the impulse response indicate multi-path reflections in the acoustic signal and give some indication of acoustic path lengths.

12.3.2.2 *Interpretation of Impulse Response for a 2CH Adaptive Filter*

The impulse response is most effectively used in the 2CH adaptive filter mode. Information derivable from the impulse response includes the proper delay adjustments, the degree of filter convergence, and filter size requirements.

Input delay adjustment is generally a non-critical operation; however, precise adjustment of this parameter permits more efficient utilization of the adaptive filter.

Current adaptive filter designs are so large that, for most cases, the delay may be set to a fixed value, allowing the adaptive filter to automatically provide the required delay. A conservative procedure is to set the *primary* delay to at least 5–6 milliseconds to assure proper signal registration in the filter. The reference delay should be set to zero. The following discussion may be helpful in understanding the delay feature as well as in assisting the operator in making use of the delays to compensate for long acoustic paths.

Consider the configuration of Figure 12-5. In this figure, the noise source along with the primary and reference microphones is shown. The adaptive filter in the 2CH mode is made up of three functions: two adjustable input delays, an adaptive filter, and a differencer. In practice, only a single delay line is used, and it is switched into either input path. The desired signal is not shown since only the noise is considered in delay adjustments.

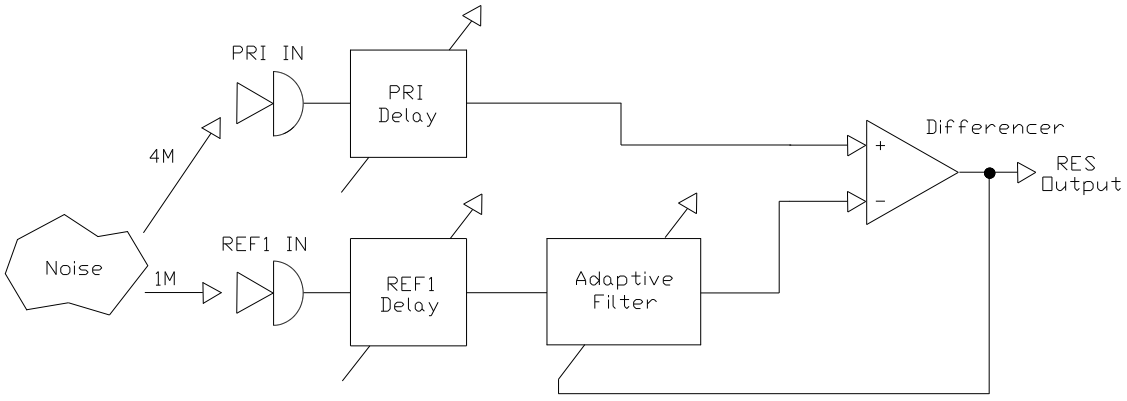


Figure 12-5: Adaptive Noise Cancellation Delay Example

In this example, the primary delay is set to zero, and the reference delay is adjusted to remove excess acoustic path delay. In the figure, the noise travels 1 meter to reach the reference microphone and 4 meters to reach the primary microphone. As sound travels at approximately 1 meter per 3 msec (1140 ft/s), the direct path delay difference between the noise at the two microphones is

$$4\text{ m} - 1\text{ m} = 3\text{ m},$$

or in milliseconds

12 msec – 3 msec = 9 msec.

This is for the direct path. The reference delay is adjusted to reduce the direct path difference yet still leave a *slightly longer primary path* than reference path. Reflected paths are always longer and are not considered in delay adjustments.

If no delays were present in the system shown, the noise signal would reach the adaptive filter 9 msec before the primary signal reached the differencer. The adaptive filter would then automatically introduce a 9 msec delay to adjust for this fixed direct path difference. In this case, the filter would automatically set the first 9 msec of coefficients to zero accomplishing this. The leading 9 msec of the displayed impulse response would thus be of small magnitude. Using the adaptive filter as a delay line is an inefficient use of small adaptive filters.

Note that if the primary delay were nonzero, the delay difference would actually increase, thus causing the adaptive filter to introduce more delay and leaving less filter to cancel reflected paths. To increase the utilization efficiency of the adaptive filter, in this case the REF DELAY should be increased from zero to a value less than 9 msec.

In Figure 12-5, the reference delay might be adjusted to reduce the acoustic delay difference from 9 msec to 3 msec. This delay difference should have a minimal positive value to allow the reference signal to reach the adaptive filter *before* the primary signal reaches the differencer. The impulse response would thus be shifted left by 6 msec on the oscilloscope. If the reference delay were increased more than 9 msec, the direct path, which normally carries the most energy, would be completely missed. Too much delay in the reference path drastically reduces the cancellation capability of the system. Too much delay in the primary path is normally not detrimental since current generation adaptive filters are quite large and may be used somewhat inefficiently without penalty.

The coefficient display moves to the left (less delay difference) when reference delay is increased and to the right (more delay difference) when primary delay is increased. The input delay is properly adjusted when the coefficient display starts on the left with a short, small-magnitude tail, is followed by a region of large magnitude, and then trails off toward zero.

A minimal primary path delay of 3 to 5 msec is normally required for proper operation of the adaptive filter. The reference delay should never be adjusted to remove this entire path, as noise cancellation would be substantially limited. The reason is mathematical in nature and beyond the scope of this manual.

Figure 12-6 is a hypothetical coefficient display for the example. The display starts with 3 msec of small magnitude coefficients, corresponding to the required 3 msec of excess primary path delay. Next is a region of large coefficients corresponding to the direct path cancellation by the adaptive filter. A strong reflected path is evidenced by the peak at approximately 15 msec, and a second significant-energy reflected path is indicated at approximately 26 msec. Acoustic reflections are often dispersed in time, causing peaks in the impulse response to also be spread.

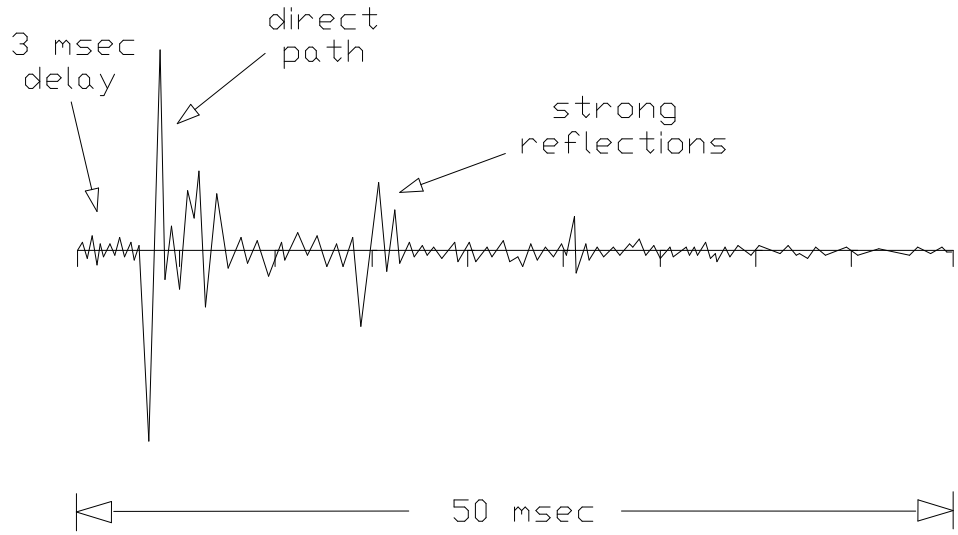


Figure 12-6: Hypothetical Oscilloscope Impulse Display for Example with 6 msec Delay in Reference Path

If the primary input gain is reduced while keeping the reference level constant, the filter coefficients will preserve the same pattern but decrease in magnitude. The gain of the adaptive filter, which is controlled by the magnitude of the impulse response, must exactly match that of the primary signal to achieve maximum cancellation.

Some DAC filters (such as the MicroDAC IV) incorporate an automatic reference compensation algorithm. This feature automatically compensates for maladjustments between the primary and reference signal input levels, thus removing a major critical adjustment in instrument operation. If the reference level is reduced significantly, the coefficients will grow to a maximum value but then will slowly reduce themselves as this algorithm readjusts the reference level. This feature is extremely useful when the filter needs to be set up and operated by non-technical personnel.

The coefficient display also illustrates the dynamics of the convergence process. Pressing the filter clear pushbutton zeros the coefficients, and the filter begins the convergence process anew. The 2CH convergence times are generally much longer than 1CH times due to the smaller degree of signal correlation. The major portion of convergence takes place in the first few seconds; however, very small changes in the impulse response can be observed over longer periods.

EXERCISES

1. Strong spectral lines on an analog recording are observed at 185, 370, and 555 Hz. The tape player is calibrated and is playing back at precisely 3.75 ips. At what speed was the tape recorded?

13. ELECTRONIC SURVEILLANCE AND COUNTERMEASURES

Electronic surveillance (ES) has been used for decades, but it has become increasingly common in today's technically advanced society. As science and technology advance, the threat of technical penetration increases. Modern technology allows advanced electronic aids and eavesdropping devices to extend of the limits of hearing. On the other hand, technical countermeasures capabilities have also expanded, and knowledge of both surveillance techniques and audio processing can combine to aid in the prevention and detection of eavesdropping. Audio surveillance and counter-surveillance are constantly competing in an attempt to stay one step ahead of each other.

This chapter covers the technical details relevant to electronic eavesdropping, but it is not intended to convey how to conduct surveillance or counter-surveillance. In addition to knowing the technical facts about electronic surveillance, various other human factors and practical procedures (*i.e.*, technical tradecraft) should be understood to conduct surveillance or sweeps properly.

13.1 Electronic Surveillance Overview

Electronic surveillance is a constantly advancing field. Amateurs and hobbyists can easily build surveillance devices that were the state of the art a decade or two ago. "Spy shops" sell various surveillance and counter-surveillance equipment to the public. Meanwhile, the threshold of electronic surveillance is continually being pushed as new devices pack greater capabilities into a smaller space with less power consumption. In competition with technical surveillance, the field of technical surveillance countermeasures (TSCM) has grown. A good countermeasures expert knows both sides of the game.

The threat of electronic surveillance can come from hostile intelligence services (HOIS), criminal activity, industrial espionage (IE), and internal eavesdropping. Government agencies specially train countermeasures technicians who can conduct TSCM surveys and inspections. Some marginally skilled professionals sell their services in using their equipment to perform a "bug sweep" for a large fee, but proper training in TSCM involves skills in investigative, electronic, and physical security.

Although there are a number of good-quality, legitimate TSCM devices available, some supposed countermeasures equipment has little function more than blinking lights. Also, with the advancement of technology, today's inexpensive countermeasures equipment can often detect and defeat only yesterday's bugs. The most recent capabilities are needed for countermeasures if the surveillance equipment is also the state of the art.

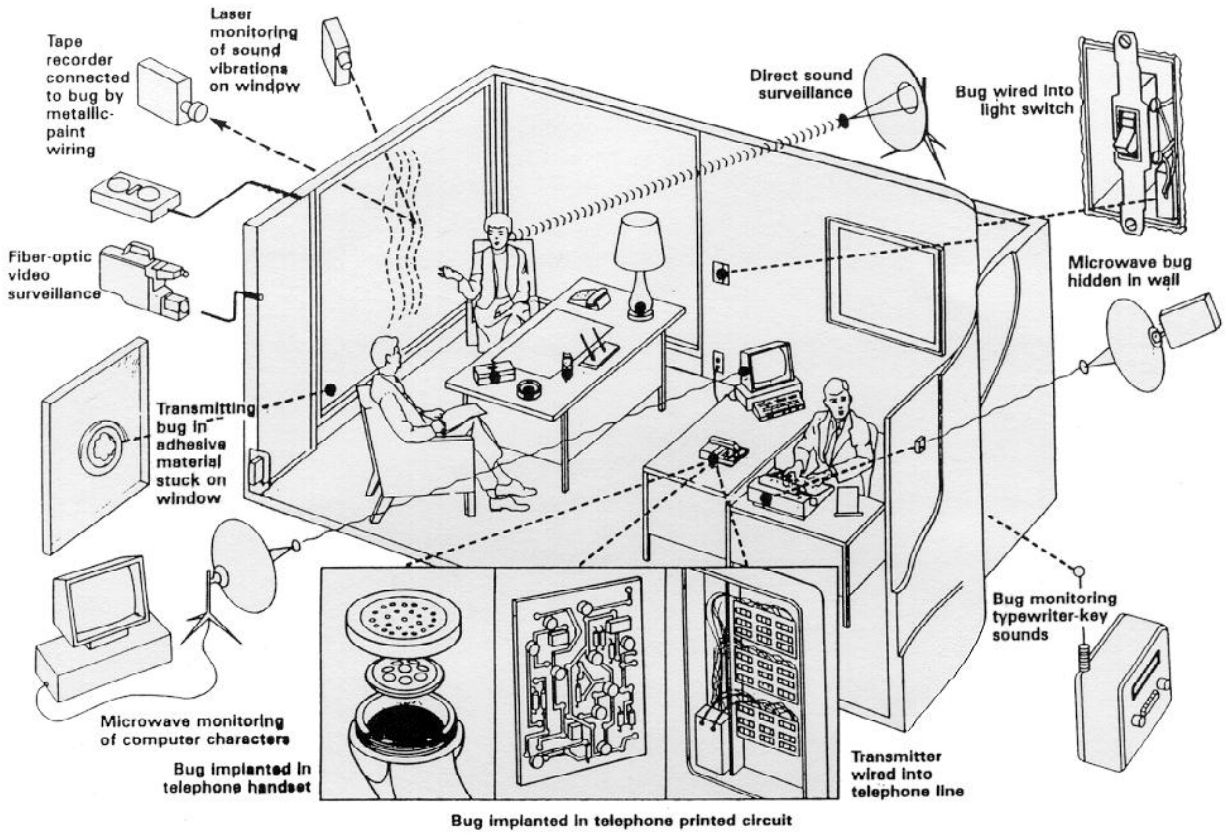


Figure 13-1: Overview of Surveillance Devices

13.2 Audio Surveillance

A review of technical surveillance is warranted before entering into details of countermeasures. An understanding of the systems that we are trying to discover and defeat is necessary for success in TSCM. Many of the techniques for audio surveillance rely upon microphones, recorders, and an understanding of the acoustic environment, as related in Chapters 6, 7, and 8. The most important qualities of surveillance devices are how they are powered, how they receive an audio signal, and how they transmit the signal.

13.2.1 Power

Surveillance devices must be powered either by a battery or by leeching power from an in-place power source.

Battery-powered devices have a limited lifetime, often days to weeks, though some are voice-activated (see Section 13.2.2.1) to extend their usable lifetime. Devices running off battery power also attempt to draw minimum current and hence have more limited capabilities.

To avoid the need for batteries, many bug-transmitters steal power from existing sources such as telephone lines or AC power lines. However, they must be planted such that they are always "plugged in." These devices are often called "leeches," "suckers," or "parasites." Some devices can be planted inside an electronic product such as a television set or radio to draw power though the electronic device. AC-powered devices have access to practically as much power as desired, though devices powered from a telephone line may actually have less current available than a battery-powered device.

Devices that use AC power need to drop the mains voltage level (100–240 V AC) to a low voltage (around 6–18 V DC). This transformation requires extra circuitry. For safety reasons, devices attached directly to power lines require that the power be disconnected from the immediate location for the duration of the installation. Similarly, any unit should be isolated from the mains supply before it is manipulated or removed.

Units that draw power from the telephone line are easier to detect with measurements of electricity. If too much current is drawn, the line will be dragged down to register permanently in the off-hook condition.

Some models of transmitters receive their primary power from a mains or telephone voltage supply but also have batteries to cover the loss of mains power. These devices may even use the mains or telephone power to constantly charge the rechargeable batteries.

Condenser microphones and amplifier units may be present on a wire pair but become active only when power is supplied along the wires. Providing *phantom power* to such a device activates it and enables its detection. Such devices typically need 1 to 52 V DC. In a recording studio, such phantom power comes from the wires that connect the mic to a recorder, sound amplifier, or audio console; in surveillance, the power may come from an existing source or may be remotely supplied to control the device.

13.2.2 Receiving Information

Collecting the sound data is the important first step in audio surveillance. The most popular forms of collection are taps and bugs, but there are also induction devices and a large number of techniques for collecting data from telephones. A number of systems, regardless of the collection method, incorporate voice-actuated switches.

13.2.2.1 Voice-Actuated Switches

A voice-actuated switch—also known as an audio actuator, voice-operated relay (VOR), sound operated relay (SOR), or voice-operated transmit (VOX)—is an electronic device that actuates another device when it detects audio frequency energy. It is commonly used for the sake of

conserving recording media and/or power consumption when there is no audio signal present to record or transmit. To conserve battery power, many bug-transmitters are activated only when the conversation is in progress.

13.2.2.2 Taps

Taps involve the interception of information on communication lines, most often telephone lines, while the line is being used. The information intercepted could be voice, facsimile data, or computer modem data. Hardwire taps can be made at any point between the telephone or other communications instrument and the telephone exchange.

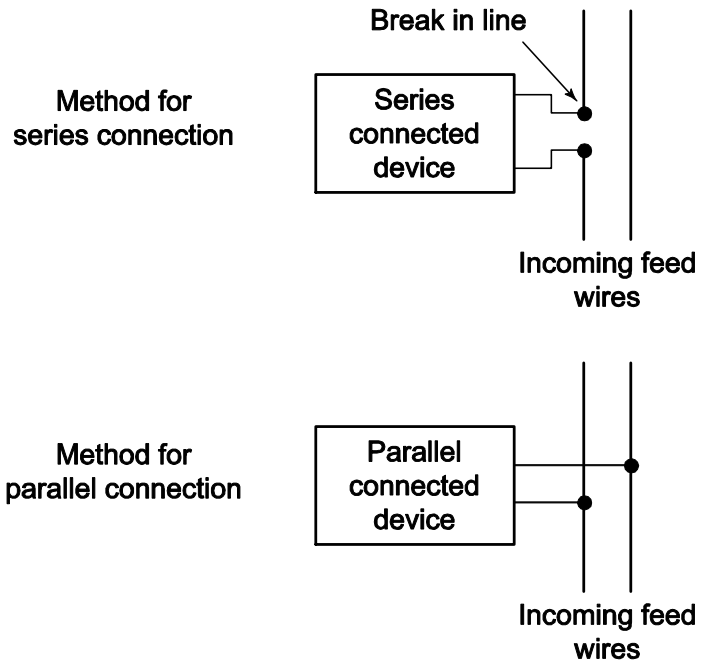


Figure 13-2: Series & Parallel Connections

A metallic tap is a pair of wires attached to the target pair in a process called jumping or bridging. The bridge can be made within a telephone set or anywhere along the wire run. An expert tapper will use filters to block the hum. A matching transformer, or matchbox, will isolate the recorder from the line. For a telephone tap, a capacitor is also used to block the DC power supplied on the telephone line.

A *parallel tap* is connected across the communications line. It often includes a radio transmitter to send the intercepted signals to a remote location. Parallel taps need high impedence to minimize the effect on the line. A *series tap*, or *parasitic tap*, is connected in series with a telephone line and draws power from the current flowing in the telephone system when the target phone is in use. Series taps need low impedance to minimize the effect on the line. For example, a phone line tap might require 8 Ω or less in series or 2 kΩ or more in parallel.

A series device is generally considered to be easier to detect. It often requires that the telephone or other system be physically disconnected to allow the insertion of the device. Counter-surveillance devices can detect when a line has been temporarily disconnected, and even an additional resistance on the line can be detected if the device is connected before the line is cut.

13.2.2.3 Bugs

Planting a microphone is the oldest and most reliable method for monitoring a room conversation. The "mic and wire" system can include switching actuators and multiple microphones. Various types of microphones, including each of those described in Section 7.1, are available in any size or form for surveillance.

Microphone placement is very important, as was covered in Chapter 7. Using two properly-spaced mics will increase intelligibility. Placing two microphones in a binaural installation can allow stereo listening. If an air-filled tube is used to convey the sound from the target location to the mic, then performing acoustical impedance matching, which is similar to electrical impedance matching, will increase the detection of desired acoustic frequencies.

Typical mic placement is in the same room as the target, but the transmission of sound means that mics can be placed not just in adjacent rooms but also along other sound-conducting pathways. For example, many air ducts carry sound away from the rooms which they adjoin, and thus they can form good locations for mics. Moreover, solid objects can form good sound-conductive paths that a surveillance expert can use to his advantage. As examples, accelerometers placed on walls or beams that are adjacent to the target area can pick up sounds, and pipes can even convey sounds away from the building.

Telephone instruments can be bugged by adding components or altering wiring inside the telephone set. Manufacturing defects and corrosion can also cause telephones to pass audio through the lines, even when off-hook, and thus cause them to act exactly like a bugged phone.

Most loudspeakers can also function as microphones (and vice versa). Thus, an eavesdropper can put a tap in the line to a room's public address loudspeakers to pick up any room sound.

13.2.2.4 Induction Devices

An induction coil enables the establishment of a telephone line tap with no direct connection to the line. In such a device, an inductive coil picks up the audio signal on a telephone line by cutting across the magnetic field of the wire pair. It is typically placed near one of the telephone handsets on the line, though it can also be placed parallel to the wire pair within a couple of feet. Nothing is attached directly to the telephone handset or wire, and there is no signal loss.

Inductive tapping can be prevented by running by the wire pair through shielded cable or a shielded conduit. In addition, all junction and terminal boxes and access panels should be locked.

Similarly, an audio transformer can be used for line coupling. Such a transformer can be placed either in series or in parallel. The advantage to using a transformer is being able to maintain the desired resistance level for the sake of hiding the device while still being able to obtain an adequately strong signal.

13.2.2.5 Other Telephone Attacks

The most common method of conducting audio surveillance is to monitor telephone conversations. Thus, telephones are generally considered to be the greatest threat to technical security. Telephones are especially vulnerable because they are designed to transmit audio signals and are present in nearly every office and home. Fax machines are also major targets for eavesdroppers.

Second-rate telephone bugs and taps can be detected by paying attention to unusual sounds. Changes in the phone line during conversation can be indicative of a bug, and any popping noises, static, or sounds from an on-hook phone are certainly suspicious. Quality bugs and taps avoid making such obvious sounds.

A *hot-mic* is a microphone in a telephone that is active regardless of whether the hook switch is closed. Effectively, it is a device that converts a telephone into a bug. It picks up not just telephone conversation but all room conversation. It uses the handset mic and transmits over the phone line.

Multiline phone systems (*e.g.*, PBXs) are harder to tap than single-line phones, and bugs for such systems are often rarer and more expensive.

Cordless telephones, which use frequencies in the 46–48 MHz range, can be intercepted by an appropriate radio receiver. Spread-spectrum and digital phones divide the audio signal across multiple frequency bands, and hence they are less susceptible to such eavesdropping.

Cellular phones are prime targets for eavesdroppers. Most analog cellular telephones are not encrypted, and thus the radio link is vulnerable to eavesdropping by intercepting the radio transmission. Digital cellular telephone transmissions are scrambled, but eavesdroppers with the proper equipment may be able to unscramble them. The following list shows the common cellular transmission modes in order of easiest to most difficult to intercept:

1. Analog (simple FM transmission)
2. TDMA (digital)
3. CDMA (digital)
4. GSM (digital)

Note that digital encoding is not the same as encryption. Encoding simply means that the data is put into a special format, but anyone who knows the algorithm and has proper equipment can still reproduce the original audio from the digital code. Encrypted data is more secure, but several standard encryption techniques have been broken.

13.2.3 Transmitting Techniques

Once the sound has been picked up, the audio signal must be conveyed by either *hardwire* or *transmitter*. Both wired and wireless systems often employ some form of signal modulation. Section 7.2.3 also comments on audio transmission.

13.2.3.1 Signal Modulation

Communication systems often shift an audio signal to a high frequency carrier for transmission. Signal modulation is practically always used for wireless communication and sometimes for wired. Complex modulation schemes are often used to try to hide the transmitted signal from detection.

Most systems that use modulation, whether along a wire or wireless system, use frequency modulation (FM) rather than amplitude modulation (AM). FM does not require as much power, allows for a smaller device, and gives a cleaner signal.

There is a variation of FM that uses a sub-carrier frequency. Such a system is called FM/FM, and devices using this scheme are called "sub-audio," "ghost," or "phantom" devices. In standard FM, the audio signal modulates the carrier frequency; for example, a carrier frequency of 100 MHz might be used in the standard VHF/FM radio band. With FM/FM, the audio signal instead modulates a sub-carrier which in turn modulates the carrier; for example, the sub-carrier might operate at 60 kHz while the carrier is 100 MHz. With such a system, anyone tuning to the carrier frequency on a standard FM receiver would hear only hissing, and an additional decoder or demodulator is needed to recover the modulating audio. Sub-carrier frequencies often range from 25 kHz to 90 kHz.

Spread-spectrum systems divide the audio signal across multiple frequency bands, and thus they are more difficult to detect. Frequency hopping makes it very difficult to intercept a message: the carrier frequency changes from frequency to frequency on a rapid, programmed basis.

Another technique often used to mask a covert transmission is known as "snuggling." The idea of snuggling is to place the covert transmission adjacent to, sometimes right on top of, a "normal" transmission, such as a nearby radio station. This can work not only for wireless transmissions but also for wired transmissions, since RF energy from commercial AM radio stations is often present on wiring.

13.2.3.2 Wired Transmission

For a hardwired transmission, a wire physically connects the eavesdropping device to the amplifier or recorder. The typical connection is via copper electrical wire, but it may also be through cable or fiber optic media. Wired bugs have several advantages over wireless bug-transmitters. The fidelity of the signal tends to be higher, the stability tends to be greater, and the pick-up radius is wider. Also, there is no need to replace batteries. Finally, radio locators can not detect the bug.

A hardwired system that uses its own specially placed set of wires does not need external power and does not necessarily radiate radio waves. If the device is located, however, then the wires can be followed right to their source. Fiber optic cable does not generate electrical disturbances and thus is less vulnerable to detection than electrical cables.

Eavesdroppers often borrow existing wires to transmit a signal. Commonly used lines include the telephone pair, AC power lines, intercom lines, security alarm wires, doorbell wires, public address system, and furnace control system.

To avoid detection, wired transmissions often use low-power signals. When an existing line is borrowed, it is especially important to bury the surveillance signal underneath the much stronger signal that is normally present on the line.

Metallic paint is an interesting alternative to metal wires. This special "paint" consists of powdered copper and an ethyl-compound adhesive. It will properly conduct low voltages, and it can even be concealed by covering it with dirt and grime or a normal non-lead paint.

Infinity transmitters, also known as infinity bugs or harmonica bugs, use the telephone line as a transmission medium. This eavesdropping device earns its name since it can theoretically be used to monitor room conversations from anywhere in the world. The infinity transmitter must be connected across a phone line in or near the telephone instrument. The bug responds to an incoming telephone call if a certain tone signal or combination of tones is applied at the time that the ringing signal is received. To use the device, the eavesdropper dials the phone number and generates a tone of a certain frequency, thereby activating the relay decoder circuit. When activated, the infinity bug effectively lifts the handset, which prevents the phone from ringing. The bug then activates the handset mic and establishes a connection. It amplifies room sounds in its vicinity, and it transmits the audio across the telephone line to which it is connected. The device's nickname of the "harmonica bug" is earned since it was originally activated by playing the specified frequency on a harmonica. Since TSCM devices can now send out sweeping tones in attempts to activate infinity bugs, later models are activated by a combination of two simultaneous tones of specific frequencies. Before the days of digital exchanges and tone dialing, these devices could activate before the target telephone rang even once, but newer telephone systems tend to ring before the infinity transmitter can answer. Electronic switching (ESS) temporarily foiled infinity bugs, but clever surveillance designers created infinity transmitters that would overcome the ESS difficulties.

A variation on the infinity bug is to have a tape recorder or transmitter that is attached to a telephone line; the audio signal is sent along the phone line only when activated by a tone. This system works similarly to the remote operation of a telephone answering machine.

13.2.3.3 Wired Wireless Transmission

In addition to transmitting along existing wires, existing wires can also be used for wireless transmissions. For example, an unused telephone line can serve as antenna. For optimal performance, cut the wire at a length appropriate for the desired broadcast frequency.

Also, existing wires can be used as transmission wires for a carrier current or "wired wireless" system. Carrier current transmitters do not transmit their signals in free space but instead use very low frequency, low power signals that flow along conducting wires. These devices use telephone wires, power lines, and other available conductors to transmit their signals. A mic is used to modulate a small oscillator, which is attached to the line via capacitors. The RF signal travels along the line and can be received elsewhere on the same line. On an AC power line, the signal can be received at any location "upstream" towards the power company. The signal is difficult to detect because of its low frequency (often 10–500 kHz).

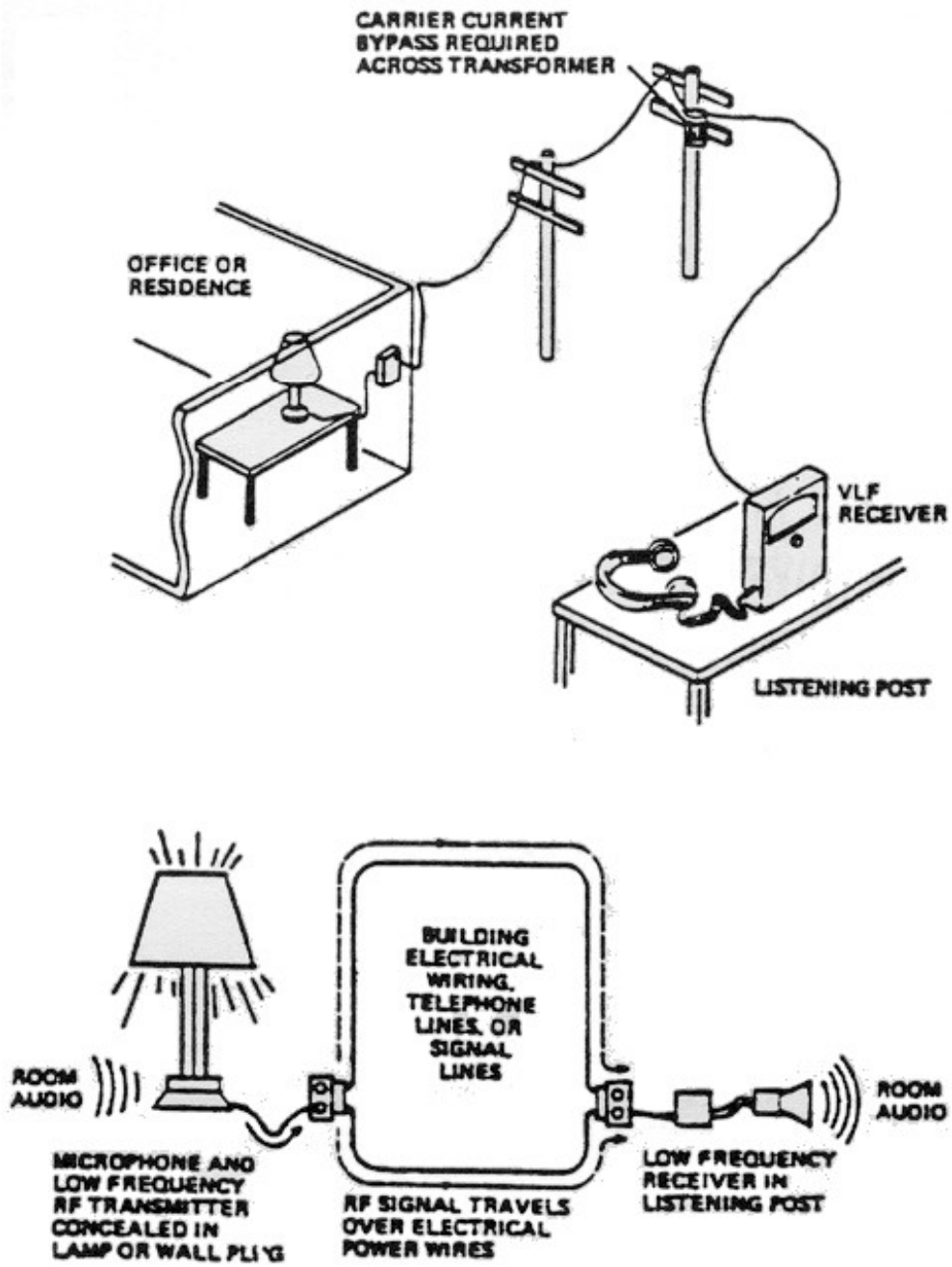


Figure 13-3: Carrier Current

13.2.3.4 Radio Frequency Transmission

Radio frequency (RF) wireless transmitters broadcast an electromagnetic signal through free space. These transmitters use techniques similar to commercial radio broadcasting but with less

power and a shorter range. RF transmission is risky since it can be detected, and therefore some form of modulation is often used to mask the transmission of the data (see Section 13.2.3.1).

Transmitters all require some form of power supply. This power can be from a battery or from stealing power from an existing telephone or AC power line.

One special form of RF transmission is from a device known as a *passive reflector* or *passive cavity transmitter*. The complete device consists of a receiver, modulator, and transmitter. A radio signal sent by the eavesdropper is transmitted to the device's reflecting capsule. This capsule, which is often a metal tube, acts as a resonator at a certain RF frequency; it also acts as a diaphragm that vibrates slightly from the acoustic waves in the room. The net result is that the capsule acts as an RF reflector which rebroadcasts the RF signal modulated by the room sound. An optional wire antenna may increase the broadcast efficiency. The device is entirely passive; it does not draw power and sends an RF signal only when one is received. Thus, it is difficult to detect except by the incoming RF energy used to activate the device.

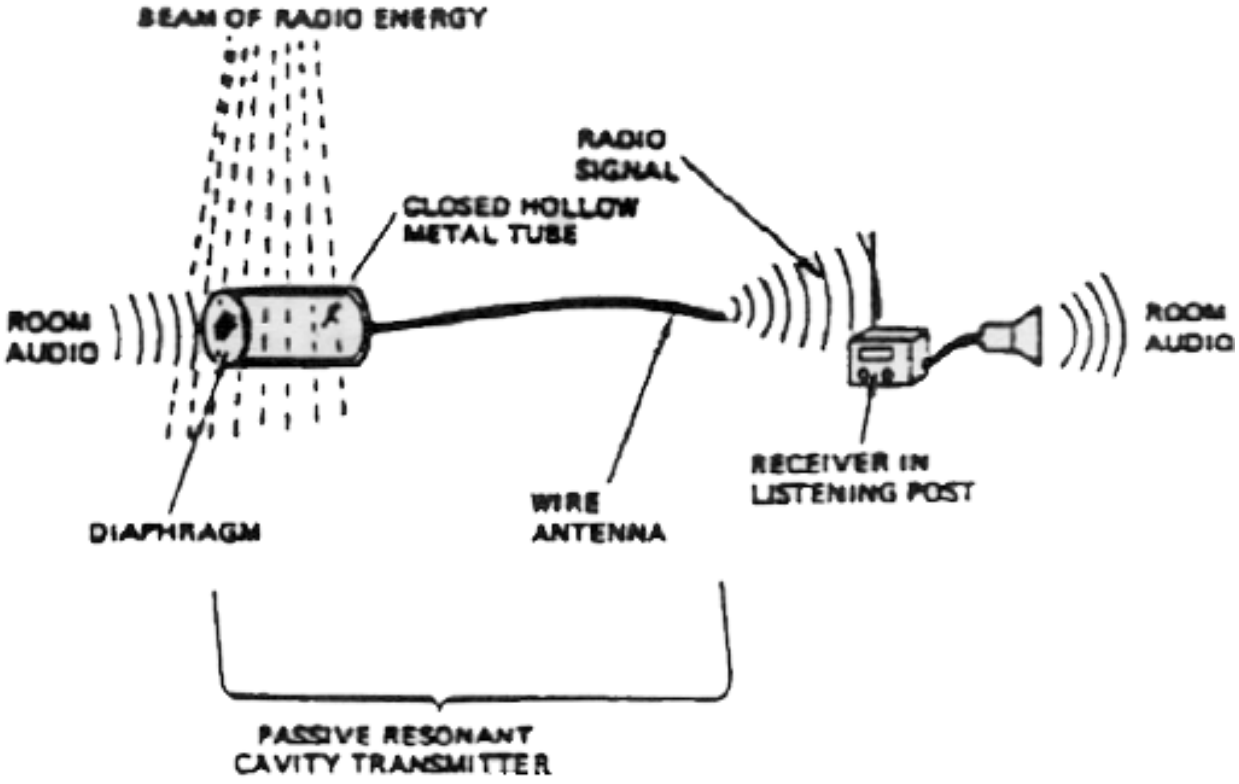


Figure 13-4: Passive Cavity Transmitter

13.2.3.5 Optical Audio Transmission

Light beams can be used as communications links. A transmitter in the target room transforms the room sound into modulated light. Often this light is in the infrared (IR) portion of the spectrum and thus is not visible. The light beam travels from the transmitter to the listening post either directly in a straight line or indirectly via reflective surfaces. In practice, however, light beam transmitters have met with only limited success. These systems are somewhat limited in range and are susceptible to interference from other infrared sources, including heat sources. Nonetheless, IR transmitters are becoming more popular, and there are now commercially-available systems which can transmit both audio and video over an IR link.

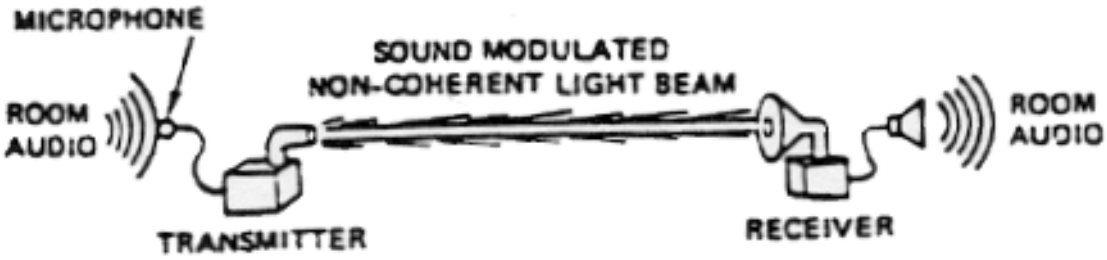


Figure 13-5: Light Beam Communications Link

13.2.3.6 Ultrasound Transmission

Ultrasonic vibrations are above the range of human hearing. Ultrasound is typically considered to be above 20,000 Hz. Although humans can not hear frequencies at that range, a number of other creatures—including insects, dogs, and dolphins—can. Ultrasound is used for sonar, medical imaging, and industrial materials testing.

Ultrasound can also be used to transmit signals. The audio signal is modulated onto a high ultrasonic carrier frequency. This carrier may be transmitted to a receiver location that is outside the room or facility under surveillance. At the receiver, the transmitted signal is demodulated and the audio or intelligence recovered. Ultrasound microphones can be very small, though the amplifiers are of moderate size.

13.3 Technical Surveillance Countermeasures (TSCM)

Countermeasures include both the prevention and detection of surveillance. A TSCM search, also known as de-bugging or sweeping, attempts to detect the presence and location of both passive and active eavesdropping and surveillance systems. Experienced specialists conduct both physical and electronic searches using technical detection equipment and methods. Because electronic monitoring is both active and passive and conducted at a distance from the target, it is difficult to uncover unless an electronic listening device is found. Countermeasures agents are

often called upon to detect the presence of listening devices that may be hidden in a room. Many devices can aid in the detection of RF bugs, telephone taps, hidden microphones, and carrier-current devices.

13.3.1 Preventing Eavesdropping

Sometimes it is easier to confuse or block a bug or tap than to find and disable it. There are several forms of masking, scrambling, and cloaking devices.

A *voice changer* does not encrypt the content of the message but changes the sound of the voice to make speaker recognition difficult.

Voice scramblers encrypt or otherwise scramble the signal prior to transmission. For a scrambler system to work, all parties engaged in the conversation must be equipped with both a scrambler and a descrambler. Most scramblers manipulate the audio signal in the frequency domain; for example, inverting scramblers switch the high and low portions of the frequency spectrum. Time-division scrambling divides the time-domain signal into short time segments and scrambles the order of transmission of the segments. Other scramblers employ masking by interjecting noise into the audio and filtering it at the receiver. Some digital scramblers operate by using a proprietary, non-standard encoding system; anyone picking up the transmission would presumably not know the proper scheme for converting the digital bit stream back into an analog waveform. Many of the simpler scrambler systems leave a fair amount of residual intelligibility, but the more sophisticated models are difficult to defeat.

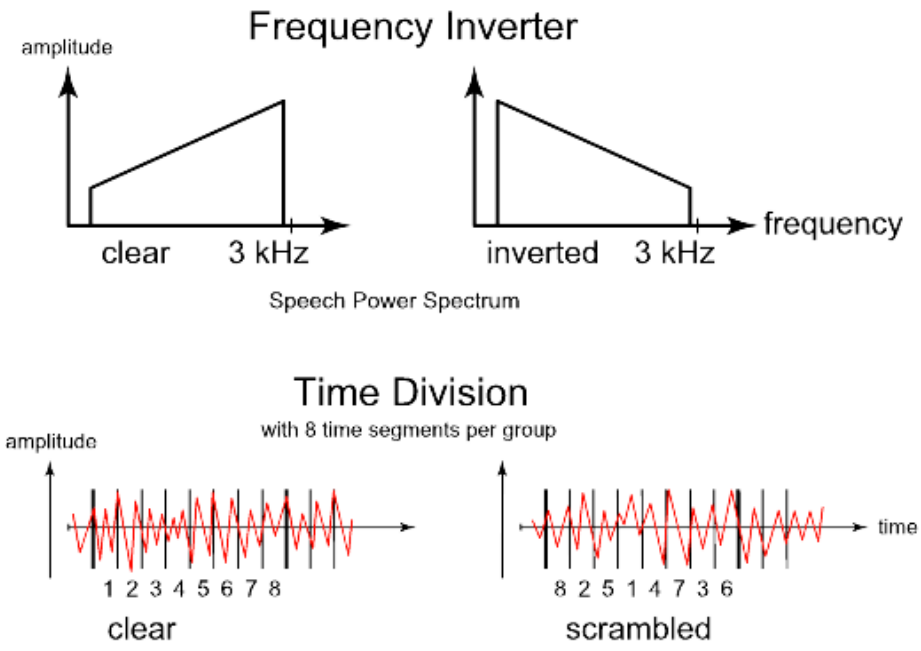


Figure 13-6: Example Types of Voice Scrambling

White noise generators produce broadband noise with the intention of masking any speech present in the room. Similar noise can also be created by turning on a shower or TV. However, appropriate audio enhancement can overcome this.

Acoustic jammers interfere with the pickup of eavesdropping devices. They generate a signal that covers range of frequencies that are audible to humans. This audio broadcast muddles the vibrations in the area and disturbs any monitoring audio receivers.

RF jammers are used to interfere with the reception of radio signals, but they generally do not blanket all frequencies at once.

Ultrasound has been used to disrupt voice-frequency microphones with mixed results. The effectiveness of the technique depends upon the type of microphone.

High-frequency *radio waves* can temporarily or even permanently disrupt nearby electronic devices, though this system defeats benign as well as benevolent devices.

Telephone conversations may be covered by taking advantage of system bandwidths. The telephone system has a bandwidth of around 3 kHz, while most tape recorders have a higher response. If a high-amplitude audio signal is generated at around 5 kHz and transmitted on the line, then it should be possible to carry out the conversation, but a tape recorder will be swamped with the high-frequency tone. Constantly broadcasting the masking signal may also wear out the power on a device powered only by batteries. An educated eavesdropper, however, could use a notch, bandpass, or lowpass filter to eliminate the masking tone.

Hot mics can be stopped by adding contacts to the bottom of the stacked switches in a telephone set so that a short circuit is formed when the instrument is hung up. Some phone taps ("drop-out relays") activate when the voltage drops to the off-hook level; these devices can be defeated by adding a circuit to load the line and change the off-hook voltage so that the tap thinks that the line is still on hook.

Carrier current transmissions can also be prevented. A high-quality surge protector or power line conditioner will often include a filter that blocks carrier current emissions; it should block RFI down to at least 1000 Hz. Shunt capacitors or lowpass filters can also be used to negate carrier current devices while still allowing the intended signal or power to flow along the circuit.

13.3.2 Detecting Eavesdropping

The following capabilities are generally useful in TSCM devices:

- Multi-meter (combination ammeter, voltmeter, and ohmmeter)
- Battery power: AC power may not be available in some locations, and it may be inconvenient to use AC power when moving around.
- High gain amplification (up to 100 dB gain)
- Highpass filter to remove 50/60-cycle hum from power lines
- Phantom power supply
- LED or similar visual indicator
- Analog or digital display meter
- Audio output so that monitored signals can be heard
- Audio amplifier
- Very high impedance for use on high-voltage power mains

A multi-meter, or volt-ohm meter (VOM), is one of the most important pieces of TSCM equipment. Section 13.3.2.2 below details use of a multi-meter on a telephone line. A multi-meter can find many taps, but modern devices can use high-input resistance, induction, and capacitance to avoid showing any noticeable change. DAC's ProbeAMP has a digital multi-meter built in.

13.3.2.1 Radio Frequency Detectors

There are several examples of practical RF detectors. Frequency counters and portable scanners can scan through a wide range of frequencies to seek out wireless transmitters. A good RF detector should cover carrier frequencies from 1 kHz to over 1 GHz. Various modulation schemes should be checked, especially since a devious eavesdropper can use any number of communications methods to try to hide the signal. Standard systems include AM, FM, sub-carrier, carrier-only, single sideband (SSB), and single sideband suppressed carrier (SSBSC). Professional systems can also detect scrambled and spread-spectrum transmissions.

FM produces less static and noise than AM. RF probes are available for picking up signals from 100 kHz to 100 MHz. The output from the probe is input to a field strength meter and/or speaker. These systems vary in sensitivity but are for short-range work only. Picking up harmonics of an RF signal, which is the same concept as harmonics of an audio signal as described in Section **Error! Reference source not found.**, is indicative of a bug since the U.S. FCC requires by law that broadcasters use equipment to eliminate harmonics.

Field strength meters give an indication that an RF signal is being transmitted, and they also show the relative field strength. Wide bandwidth is good for counter-surveillance since the

transmission frequency is unknown. A basic field strength meter is sometimes adequate, but the detection of very low powered transmitters often requires a more sensitive device.

RF detectors can indicate the presence of a signal but do not reveal its contents. The only way to confirm the origin of the signal is to listen to the unmodulated signal, which generally requires using a receiver

A common yet simple method for bug detection is to use a *squealer*. This device consists of a tone generator and RF receiver. The squealer emits an audio tone and allows it to feed back if received internally.

Most radio receivers will re-radiate radio signals, called "birdies," themselves while in operation. An eavesdropper can detect the radiation from a sweep receiver in order to know when to deactivate his bugs. Some receivers radiate a stronger signal than others, so it is preferable to use a receiver which is more difficult to detect.

13.3.2.2 Telephone Line Analysis

Several good techniques comprise telephone line analysis. Test for room sound signals and carrier current signals carrying room sounds. Perform tests at a point where the lines appear on terminals. The proper equipment can allow the test for carrier current signals to be done without making an electrical connection to the lines. Tests for audio frequency signal should be done with an audio amplifier connected directly to the line with a blocking capacitor, but a capacitive pickup device could be used instead. A multi-meter should be used to measure voltage, current, and impedance.

Voltage measurements are useful. In the on-hook state, the normal voltage across the line wires is around 48–51 V with no current flowing. In the off-hook state, the voltage falls to around 5–12 V and current does flow. The actual voltages for the on-hook and off-hook modes vary according to the telephone handset, other telephone sets on the same line, and the telephone system. If a there is a series tap on the line, then the on-hook voltage will be unaffected, but the off-hook voltage will be less than if there were no tap. There will be no voltage drop across the tap when the telephone is on the hook, but there will be a voltage drop when it is off the hook. Series taps strive for minimal resistance since it will (by Ohm's law: $V = I \times R$) minimize the voltage drop. In comparison, a device connected in parallel will drive the voltage down slightly in both the on-hook and off-hook states. Parallel taps strive for maximal resistance on the order of megohms.

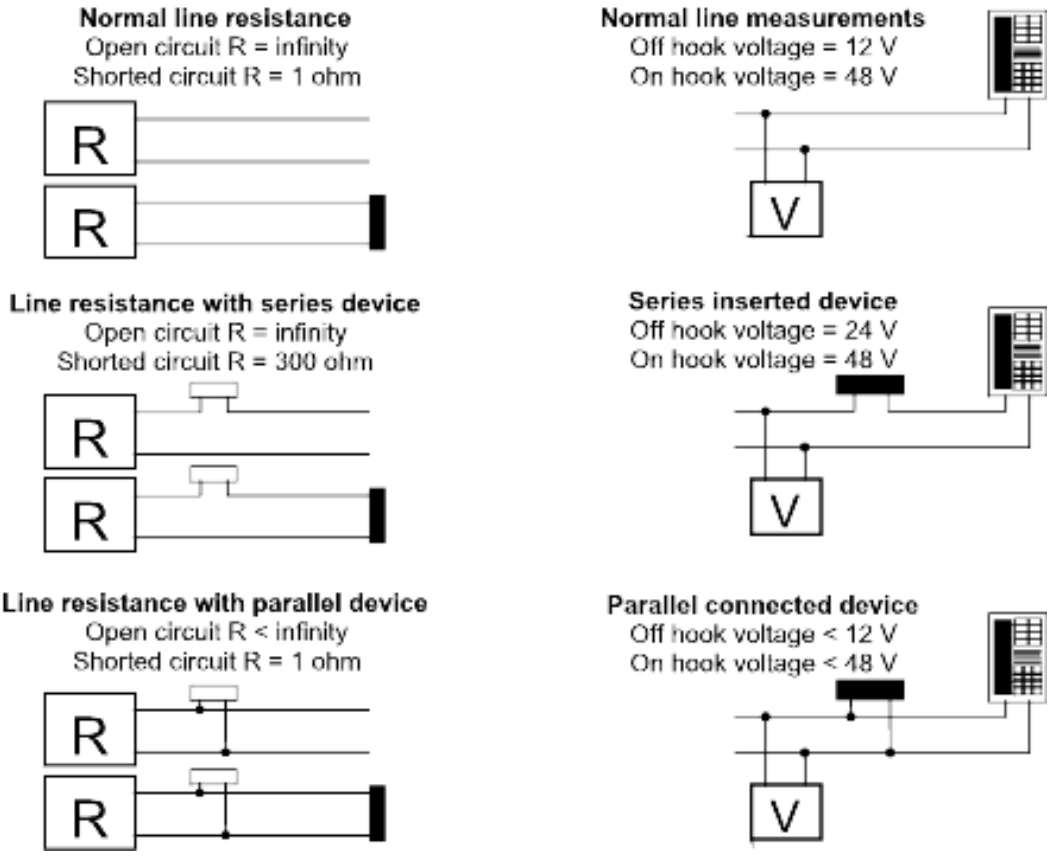


Figure 13-7: Voltage & Resistance Measurements on a Telephone Line

Resistance measurements are also useful. If the resistance is measured with all telephone sets and other equipment off-line, it should show a resistance of many megohms ("infinite" resistance). With the ends of the line pair shorted at the farthest point possible, the resistance should be no more than a couple of ohms. With a series device placed on the line, then the open-circuit resistance would still be infinite, but the short-circuit resistance would be much greater, perhaps on the order of hundreds of ohms. With a parallel tap, the open-circuit resistance would be less than infinite, and the short-circuit resistance would still be less than one ohm.

Line impedance tests send an audio frequency signal down the line, which is shorted at the far end. A load resistance and oscilloscope are placed in series with the line. The audio signal generator sweeps frequencies from a few hertz up to several kilohertz, and the oscilloscope is used to measure the voltage across the load resistor. If a device is connected in parallel on the line, then the voltage will drop across the load resistor when the generated audio frequency matches the resonant frequency of the audio components of the parallel device.

A Time-Domain Reflectometer (TDR) can be quite useful in locating taps on phone lines. A TDR uses the concepts of voltage reflection on transmission lines. When a voltage wave traveling down a transmission line encounters a load—whether from a termination, tap, or change in the line's electrical properties—a reflected wave is sent back towards the source.

Thus, a TDR sends a series of pulses down the line and notes the timing and amplitude of any reflected pulses. Analysis of these reflections can reveal the location of any loads on the line, and hence a map can be made of every electrical junction on the circuit. The line should be initially checked with a TDR when it is known to be clean, and a comparison check later can indicate whether any changes have been made to the line. This theory of TDRs applies to all transmission lines, but it is applied to telephone lines more often than AC power lines. Despite the usefulness of the device, the signal from the TDR is detectable by the eavesdropper if he is listening for it.

It is also common to use an RF signal for telephone line tracing. However, an eavesdropper may watch the tap line for that type of energy as an indication of sweep activity. Thus, a weak signal should be used to reduce the likelihood of detection.

Hot mics can be detected by searching for a signal on the telephone line while the phone is on the hook. An induction device can accomplish this without having to modify the phone or line.

A Telephone Line Analyzer (TLA) can perform several of the above measurements. It includes a multi-meter and often more devices. Many TLAs can also run a tone sweep to check for infinity bugs. Some also include TDRs. These systems are not infallible but are useful for security.

13.3.2.3 Other TSCM Devices

A number of other devices are also quite useful in TSCM.

Countermeasures agents can use ultrasonic microphones and dedicated hardware spectrum analyzers to observe the presence of high frequency acoustic energy. A limitation of this system is that the spectrum analyzer cannot tell the difference between a benign signal, such as that produced by a motion sensor, and a malicious one. DAC's UltraScope product can not only detect an ultrasonic signal but also demodulate it to allow the user to listen to the signal of interest.

A *nonlinear junction detector* (NLJD), also known as a PN junction detector, is a technical device intended to detect nonlinear junctions. A nonlinear junction occurs where two pieces of metal interface with oxidized metal between them. These junctions occur in diodes and transistors, including those in solid-state chips, which the NLJD is intended to locate; however, the same junction also occurs where a rusty nail is embedded in metal. The electronic components need not be active to be detected. On the other hand, the detector will not locate devices built from field effect transistors (FETs). The NLJD transmits the frequency and the receiver simultaneously searches for any signal at twice the transmitter frequency. NLJDs use a known frequency in their operation (915 MHz ± 13 MHz), so good surveillance technicians can monitor to know when one is being used so that they can immediately deactivate their bugs during the sweep.

High-quality *tape recorders* can be detected because they use a system that uses an oscillator of around 100 kHz. This bias oscillator can be detected from several feet away, though a little shielding can make it undetectable. Another approach to detecting tape recorders is to sense the DC field that the motor brushes produce, but such detectors are larger and more expensive.

Metal detectors can find metallic devices behind walls. Some mics are made of plastic or ceramic material to avoid such detection. Also, mics can be concealed near metal nails, junction boxes, and wiring, thus masking them from the metal detector.

Infrared (IR) *detectors* can be used to detect both laserbugs and light beams that are used as communications links. Broadband IR sensors tend to be more useful than narrowband sensors. For locating the source of an IR signal, a steerable array of IR sensors can be made to work similar to the microphone arrays described in Section 7.4.2. Thermal imaging cameras can also detect IR beams and produce a visible image of their paths.

X-ray equipment is costly but can be used to effectively scan likely hiding places.

Many electronic surveillance devices emit radiation that interferes with nearby devices, and such *emissions* can compromise their presence. Standards set by the FCC, EBU, and similar organizations insist on shielding, and high-quality devices will have good shielding. Many surveillance devices have inadequate shielding, however, and they induce static, buzzing, or erratic functioning when located near other electronics devices.

Carrier current devices can be detected with a TDR, oscilloscope, or carrier-current receiver. Alternatively, they can also be detected using DAC's ProbeAMP and UltraScope products connected together.

APPENDIX A

ANSWERS TO SIGNAL PROCESSING EXERCISES

Chapter 1

1. 0.025 W/ft^2, 0.00625 W/ft^2
2. 2.00 volts RMS
   2.00 volts RMS
3. 0.000775 volts RMS, 66dB
4. 100 watts
5. 46 dBA SPL, 59 dBB SPL, 65 dBC SPL
6. 40 dB, 5 V_{rms}, 16.2 dBm, 14.0 dBV
7. a) 0dBV, +2.2 dBm
   b) 7.75 V_{rms}
   c) –37.8 dBm, 10 mV_{rms}
   d) –3.8 dB

Chapter 2

1. 1 msec, 500 kHz
2. 7.8125 Hz, 15.625 Hz, 6.25 kHz
3. 2 kHz, 500 Hz

Chapter 3

1. 1250 Ω, 250.1 Ω, 100 Ω
2. 3.33 Ω, 0.99 Ω, 13.95 Ω

Chapter 4

1. fricatives
2. fricative, vocal cords (glottis)
3. ee (/i/), aw (/○/), 2290 Hz, 840 Hz
4. /bæt/, /ʃut/, /sloU/

Chapter 5

1. noise, noise, effect, noise, effect
2. highly muffled
3. 3 kHz, 4.5 kHz, 6 kHz, etc.
4. motor is predictable, airflow is random
5. Decrease, Reverberations absorbed by material, Audio would be reduced by muffling

Chapter 7

1. Electret
2. 1.13', 4.99', 16.6'
3. 18 dB

Chapter 8

1. 0.00023", 0.00047", 0.00094", No. Only 3 ¾ ips
2. Linearize BH curve; the playback head lowpass filters the signal
3. High frequencies
4. Distortion, wow and flutter, cross talk, noise floor, azimuth
5. Head cleaning

Chapter 9

1. Highpass filter, lowpass filter
2. 0 dB, 12 dB, 36 dB, 48 dB
3. 24 dB, 48 dB
4. 24 dB/octave
5. Bessel, Butterworth
6. 48 dB, 24 dB, 0 dB
7. Cutoff frequency, rolloff rate, stopband attenuation
8. 100 Hz; 950 Hz and 1050 Hz
9. 15 dB, 5 dB

Chapter 10

1. 6 kHz; approx 5 kHz, depending upon LPF rolloff
2. 48 dB
3. 204.8 msec
4. 60 Hz, bandlimit to 500 Hz
5. a) 01100111
   b) 00010001
   c) 147
6. 010110

Chapter 11

1. 221 msec, >2210 taps

Chapter 12

1. 3.65 ips

BIBLIOGRAPHY

Ballou, Glen, ed. *Handbook for Sound Engineers: The New Audio Cyclopedia*. 2nd ed. Carmel, Indiana: SAMS, 1991.

Hollien, Harry. *The Acoustics of Crime: The New Science of Forensic Phonetics*. New York: Plenum, 1990.

Koenig, Bruce E. "Enhancement of Forensic Audio Recordings," *Journal of the Audio Engineering Society*, Vol. 36, No. 11, Nov. 1988, pp. 884–94.

Moore, Brian C. J. *An Introduction to the Psychology of Hearing*. 4th ed. San Diego: Academic Press, 1997.

O'Shaughnessy, Douglas. *Speech Communication: Human and Machine*. Reading, Mass.: Addison-Wesley, 1987.

Rossing, Thomas D. *The Science of Sound*. 2nd ed. Addison-Wesley, 1990.