

Audio Processing

Audio processing covers many diverse fields, all involved in presenting sound to human listeners. Three areas are prominent: (1) *high fidelity music reproduction*, such as in audio compact discs, (2) *voice telecommunications*, another name for telephone networks, and (3) *synthetic speech*, where computers generate and recognize human voice patterns. While these applications have different goals and problems, they are linked by a common umpire: the human ear. Digital Signal Processing has produced revolutionary changes in these and other areas of audio processing.

Human Hearing

The human ear is an exceedingly complex organ. To make matters even more difficult, the information from *two* ears is combined in a perplexing neural network, the human brain. Keep in mind that the following is only a brief overview; there are many subtle effects and poorly understood phenomena related to human hearing.

Figure 22-1 illustrates the major structures and processes that comprise the human ear. The *outer ear* is composed of two parts, the visible flap of skin and cartilage attached to the side of the head, and the *ear canal*, a tube about 0.5 cm in diameter extending about 3 cm into the head. These structures direct environmental sounds to the sensitive *middle and inner ear* organs located safely inside of the skull bones. Stretched across the end of the ear canal is a thin sheet of tissue called the *tympanic membrane* or *ear drum*. Sound waves striking the tympanic membrane cause it to vibrate. The middle ear is a set of small bones that transfer this vibration to the *cochlea* (inner ear) where it is converted to neural impulses. The cochlea is a liquid filled tube roughly 2 mm in diameter and 3 cm in length. Although shown straight in Fig. 22-1, the cochlea is curled up and looks like a small snail shell. In fact, *cochlea* is derived from the Greek word for *snail*.

音频处理

音频处理涵盖众多不同领域，这些领域都涉及向人类听众呈现声音。其中三个领域尤为突出：(1)高保真音乐再现（如音频光盘），(2)语音通信（即电话网络），以及(3)合成语音（计算机生成并识别人类语音模式）。尽管这些应用目标和问题各不相同，但它们都与一个共同的裁判者——人耳——紧密相连。数字信号处理技术已在这些及其他音频处理领域引发了革命性变革。

人类听觉

人类的耳朵是一个极其复杂的器官。更困难的是，来自两只耳朵的信息在一个令人困惑的神经网络——人类大脑中结合在一起。请记住，以下只是一个简要概述；与人类听力相关的许多微妙效应和不为人知的现象。

图22-1展示了构成人耳的主要结构和过程。外耳由两部分组成：附着在头部侧面的可见皮肤和软骨瓣，以及直径约0.5厘米、延伸约3厘米进入头部的耳道。这些结构将环境声音传导至位于颅骨内部的安全区域的敏感中耳和内耳器官。横跨耳道末端的是一层称为鼓膜或耳鼓的薄组织。声波撞击鼓膜使其振动。中耳是一组小骨头，将这种振动传递至耳蜗（内耳），并在其中转化为神经冲动。耳蜗是一个直径约2毫米、长度3厘米的充满液体的管道。尽管图22-1中显示为直线结构，但耳蜗实际上是卷曲的，形似小蜗牛壳。事实上，耳蜗一词源自希腊语中表示蜗牛的词汇。

When a sound wave tries to pass from air into liquid, only a small fraction of the sound is transmitted through the interface, while the remainder of the energy is reflected. This is because air has a *low* mechanical impedance (low acoustic pressure and high particle velocity resulting from low density and high compressibility), while liquid has a *high* mechanical impedance. In less technical terms, it requires more effort to wave your hand in water than it does to wave it in air. This difference in mechanical impedance results in most of the sound being reflected at an air/liquid interface.

The middle ear is an *impedance matching* network that increases the fraction of sound energy entering the liquid of the inner ear. For example, fish do not have an ear drum or middle ear, because they have no need to hear in air. Most of the impedance conversion results from the difference in *area* between the ear drum (receiving sound from the air) and the *oval window* (transmitting sound into the liquid, see Fig. 22-1). The ear drum has an area of about 60 (mm)^2 , while the oval window has an area of roughly 4 (mm)^2 . Since pressure is equal to force divided by area, this difference in area increases the sound wave pressure by about 15 times.

Contained within the cochlea is the *basilar membrane*, the supporting structure for about 12,000 sensory cells forming the *cochlear nerve*. The basilar membrane is stiffest near the oval window, and becomes more flexible toward the opposite end, allowing it to act as a *frequency spectrum analyzer*. When exposed to a high frequency signal, the basilar membrane resonates where it is stiff, resulting in the excitation of nerve cells close to the oval window. Likewise, low frequency sounds excite nerve cells at the far end of the basilar membrane. This makes specific fibers in the cochlear nerve respond to specific frequencies. This organization is called the **place principle**, and is preserved throughout the auditory pathway into the brain.

Another information encoding scheme is also used in human hearing, called the **volley principle**. Nerve cells transmit information by generating brief electrical pulses called *action potentials*. A nerve cell on the basilar membrane can encode audio information by producing an action potential in response to each cycle of the vibration. For example, a 200 hertz sound wave can be represented by a neuron producing 200 action potentials per second. However, this only works at frequencies below about 500 hertz, the maximum rate that neurons can produce action potentials. The human ear overcomes this problem by allowing several nerve cells to take turns performing this single task. For example, a 3000 hertz tone might be represented by *ten* nerve cells alternately firing at 300 times per second. This extends the range of the volley principle to about 4 kHz, above which the place principle is exclusively used.

Table 22-1 shows the relationship between sound intensity and perceived loudness. It is common to express sound intensity on a logarithmic scale, called **decibel SPL** (Sound Power Level). On this scale, 0 dB SPL is a sound wave power of $10^{-16} \text{ watts/cm}^2$, about the weakest sound detectable by the human ear. Normal speech is at about 60 dB SPL, while painful damage to the ear occurs at about 140 dB SPL.

当声波试图从空气传入液体时，仅有极小部分声能能穿透界面，其余能量则会被反射。这是因为空气具有低机械阻抗（由低密度和高可压缩性导致的声压低和粒子运动速度快），而液体则具有高机械阻抗。用通俗的话说，就像在水中挥动手臂比在空气中挥动需要更大力气。这种机械阻抗的差异导致大部分声波在气液界面处发生反射。

中耳是一个阻抗匹配网络，能够提升进入内耳液体的声能比例。例如鱼类没有鼓膜或中耳，因为它们不需要在空气中听声音。大部分阻抗转换源于鼓膜（接收来自空气的声波）与卵圆窗（将声波传递至液体中，见图22-1）面积的差异。鼓膜面积约为60平方毫米²，而卵圆窗面积仅有约4平方毫米²。由于压力等于力除以面积，这种面积差异使声波压力增大了约15倍。

耳蜗内部包含着基底膜，这是支撑着约12,000个感觉细胞的结构，这些细胞共同构成了耳蜗神经。基底膜在卵圆窗附近最为坚硬，向另一端逐渐变软，这种特性使其能够充当频率分析器。当受到高频信号刺激时，基底膜会在坚硬区域产生共振，从而激发靠近卵圆窗的神经细胞。同样地，低频声音会刺激基底膜远端的神经细胞。这种机制使得耳蜗神经中的特定纤维能够响应特定频率。这种组织结构被称为位置原理，并在通往大脑的听觉通路中始终保持着这一特性。

人类听觉系统还采用了一种名为群发原则的信息编码机制。神经细胞通过产生称为动作电位的短暂电信号来传递信息。位于基底膜的神经细胞能够通过响应振动周期产生动作电位来编码音频信息。例如，200赫兹的声波可由神经元每秒产生200次动作电位来表征。但该机制仅适用于约500赫兹以下的频率——这是神经元产生动作电位的最高速率。人类耳朵通过让多个神经细胞轮流执行同一任务来解决这一问题。例如，3000赫兹的音调可能由十个神经细胞以每秒300次的频率交替放电来表征。这种机制将群发原则的应用范围扩展至约4千赫兹，超过该频率后则完全采用位置原则。

表22-1展示了声强与感知响度之间的关系。通常采用对数刻度来表示声强，称为分贝声压级（Sound Power Level）。在此刻度下，0 dB SPL对应声波功率为 10^{-16} 瓦/平方厘米²，约为人耳可检测到的最弱声音。正常人说话时的声压级约为60 dB SPL，而造成耳部疼痛性损伤的声压级则约为140 dB SPL。

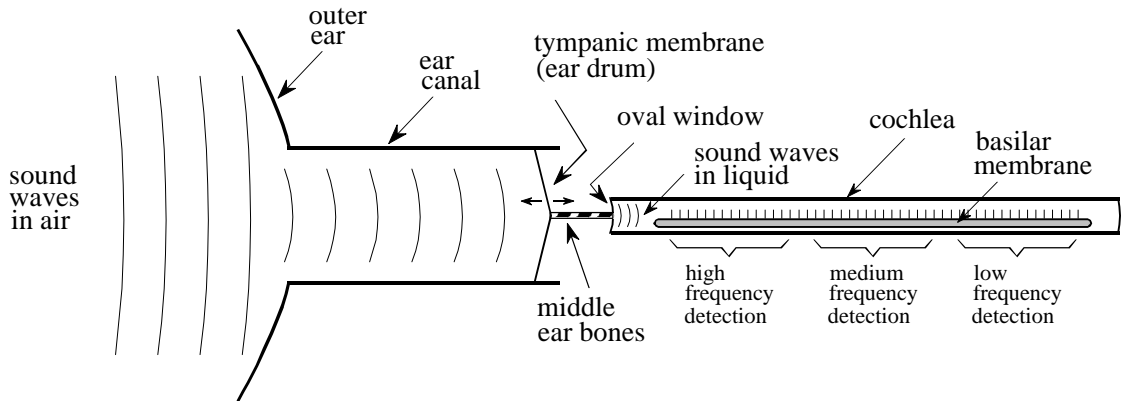


FIGURE 22-1

Functional diagram of the human ear. The outer ear collects sound waves from the environment and channels them to the tympanic membrane (ear drum), a thin sheet of tissue that vibrates in synchronization with the air waveform. The middle ear bones (hammer, anvil and stirrup) transmit these vibrations to the oval window, a flexible membrane in the fluid filled cochlea. Contained within the cochlea is the basilar membrane, the supporting structure for about 12,000 nerve cells that form the cochlear nerve. Due to the varying stiffness of the basilar membrane, each nerve cell only responds to a narrow range of audio frequencies, making the ear a frequency spectrum analyzer.

The difference between the loudest and faintest sounds that humans can hear is about 120 dB, a range of one-million in amplitude. Listeners can detect a *change* in loudness when the signal is altered by about 1 dB (a 12% change in amplitude). In other words, there are only about 120 levels of loudness that can be perceived from the faintest whisper to the loudest thunder. The sensitivity of the ear is amazing; when listening to very weak sounds, the ear drum vibrates less than the diameter of a single molecule!

The perception of loudness relates roughly to the sound power to an exponent of $1/3$. For example, if you increase the sound power by a factor of *ten*, listeners will report that the loudness has increased by a factor of about *two* ($10^{1/3} \approx 2$). This is a major problem for eliminating undesirable environmental sounds, for instance, the beefed-up stereo in the next door apartment. Suppose you diligently cover 99% of your wall with a perfect soundproof material, missing only 1% of the surface area due to doors, corners, vents, etc. Even though the sound power has been reduced to only 1% of its former value, the perceived loudness has only dropped to about $0.01^{1/3} \approx 0.2$, or 20%.

The range of human hearing is generally considered to be 20 Hz to 20 kHz, but it is far more sensitive to sounds between 1 kHz and 4 kHz. For example, listeners can detect sounds as low as 0 dB SPL at 3 kHz, but require 40 dB SPL at 100 hertz (an amplitude increase of 100). Listeners can tell that two tones are different if their frequencies differ by more than about 0.3% at 3 kHz. This increases to 3% at 100 hertz. For comparison, adjacent keys on a piano differ by about 6% in frequency.

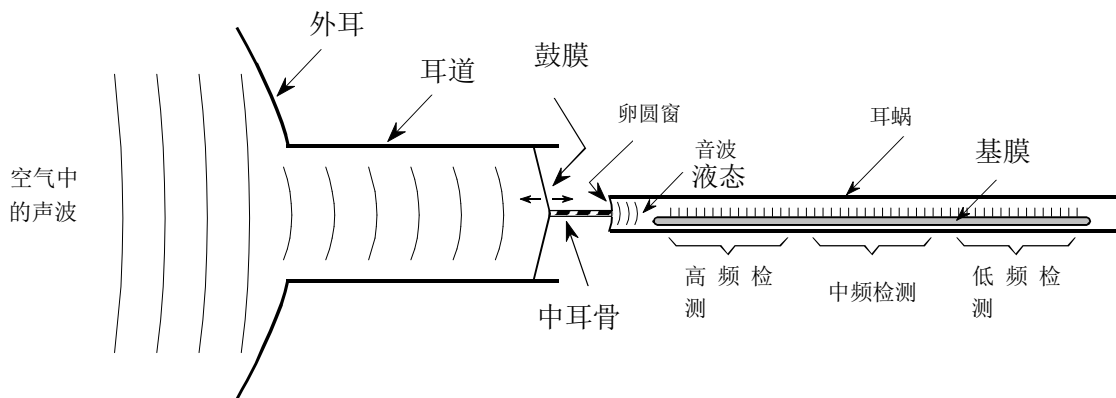


图22-1

人类耳朵的功能示意图。外耳负责收集环境中的声波，并将其传导至鼓膜（耳鼓）——这是一层薄薄的组织，会与空气波形同步振动。中耳的三块骨头（锤骨、砧骨和镫骨）将这些振动传递至卵圆窗，这是充满液体的耳蜗内的一层柔性膜。耳蜗内还包含基底膜，它是支撑约12,000个神经细胞的结构，这些神经细胞共同构成耳蜗神经。由于基底膜的硬度存在差异，每个神经细胞仅对特定范围的音频频率产生反应，这使得耳朵具备频率谱分析功能。

人类能听到的最响亮与最微弱声音之间的差异约为120分贝，相当于一百万倍的振幅变化。当信号强度改变约1分贝（相当于振幅变化12%）时，听者就能察觉到响度的变化。换句话说，从最微弱的耳语到最响亮的雷声，人类耳朵能感知的响度层次仅有约120级。人类耳朵的灵敏度令人惊叹——当听到极其微弱的声音时，耳膜振动幅度甚至小于单个分子的直径！

响度的感知大致与声功率的 $1/3$ 次方相关。例如，如果你将声功率提高十倍，听者会报告响度增加了约两倍（ $10^{1/3} \approx 2$ ）。这对于消除不希望的环境声音（例如隔壁公寓里扩音的立体声）是一个主要问题。假设你用完美的隔音材料覆盖了99%的墙面，仅因门、角落、通风口等留出1%的表面积未覆盖。即使声功率已降至原来的1%，感知响度也仅下降至约 $0.01^{1/3} \approx 0.2$ ，即20%。

人类听觉范围通常被认为在20赫兹至20千赫兹之间，但对1千赫兹至4千赫兹之间的声音更为敏感。例如，听者在3千赫兹时能检测到低至0分贝声压级的声音，但在100赫兹时需要40分贝声压级（即振幅增加100倍）。当两个音调在3千赫兹时频率差异超过约0.3%时，听者即可分辨其不同。这一阈值在100赫兹时提升至3%。作为对比，钢琴相邻键位的频率差异约为6%。

TABLE 22-1
Units of sound intensity. Sound intensity is expressed as power per unit area (such as watts/cm²), or more commonly on a logarithmic scale called *decibels SPL*. As this table shows, human hearing is the most sensitive between 1 kHz and 4 kHz.

	Watts/cm ²	Decibels SPL	Example sound
	10 ⁻²	140 dB	Pain
	10 ⁻³	130 dB	
↑	10 ⁻⁴	120 dB	Discomfort
	10 ⁻⁵	110 dB	Jack hammers and rock concerts
	10 ⁻⁶	100 dB	
	10 ⁻⁷	90 dB	OSHA limit for industrial noise
	10 ⁻⁸	80 dB	
	10 ⁻⁹	70 dB	
	10 ⁻¹⁰	60 dB	Normal conversation
	10 ⁻¹¹	50 dB	
	10 ⁻¹²	40 dB	Weakest audible at 100 hertz
	10 ⁻¹³	30 dB	
	10 ⁻¹⁴	20 dB	Weakest audible at 10kHz
	10 ⁻¹⁵	10 dB	
	10 ⁻¹⁶	0 dB	Weakest audible at 3 kHz
	10 ⁻¹⁷	-10 dB	
	10 ⁻¹⁸	-20 dB	

The primary advantage of having *two* ears is the ability to identify the *direction* of the sound. Human listeners can detect the difference between two sound sources that are placed as little as three degrees apart, about the width of a person at 10 meters. This directional information is obtained in two separate ways. First, frequencies above about 1 kHz are strongly *shadowed* by the head. In other words, the ear nearest the sound receives a stronger signal than the ear on the opposite side of the head. The second clue to directionality is that the ear on the far side of the head hears the sound slightly *later* than the near ear, due to its greater distance from the source. Based on a typical head size (about 22 cm) and the speed of sound (about 340 meters per second), an angular discrimination of three degrees requires a timing precision of about 30 microseconds. Since this timing requires the volley principle, this clue to directionality is predominately used for sounds less than about 1 kHz.

Both these sources of directional information are greatly aided by the ability to turn the head and observe the change in the signals. An interesting sensation occurs when a listener is presented with exactly the same sounds to both ears, such as listening to monaural sound through headphones. The brain concludes that the sound is coming from the center of the listener's head!

While human hearing can determine the *direction* a sound is from, it does poorly in identifying the *distance* to the sound source. This is because there are few clues available in a sound wave that can provide this information. Human hearing weakly perceives that high frequency sounds are nearby, while low frequency sounds are distant. This is because sound waves dissipate their higher frequencies as they propagate long distances. Echo content is another weak clue to distance, providing a perception of the room size. For example,

表22-1

声强的单位。声强以单位面积的功率表示（例如瓦特/平方厘米²），或者更常见的是以对数标度表示，称为分贝*SPL*。如表所示，人类在1 kHz到4 kHz之间最敏感。

	瓦/平方厘米 ²	声压级分贝	示例音频
	10 ⁻²	140分贝	疼痛
	10 ⁻³	130分贝	
	10 ⁻⁴	120分贝	不适
	10 ⁻⁵	110分贝	杰克·哈默斯和摇滚音乐会
	10 ⁻⁶	100分贝	
	10 ⁻⁷	90分贝	工业噪声的OSHA限值
	10 ⁻⁸	80分贝	
	10 ⁻⁹	70分贝	
	10 ⁻¹⁰	60分贝	正常会话
	10 ⁻¹¹	50分贝	
	10 ⁻¹²	40分贝	100赫兹时最弱可闻
	10 ⁻¹³	30分贝	
	10 ⁻¹⁴	20分贝	10kHz时最弱可听
	10 ⁻¹⁵	10分贝	
	10 ⁻¹⁶	0分贝	3 kHz时最弱可听
	10 ⁻¹⁷	-10分贝	
	10 ⁻¹⁸	-20分贝	

拥有两只耳朵的最大优势在于能够辨别声音的*方向*。人类听者只需将两个声源相距约10米（相当于一个人的宽度）的三个度角，就能分辨出它们的差异。这种方向感知主要通过两种方式实现：首先，约1千赫兹以上的高频声音会被头部*阻隔*，即靠近声源的耳朵接收到的信号比另一侧更强烈；其次，由于距离声源较远，头部另一侧的耳朵会比近侧耳朵稍晚*接收*声音。根据典型头部尺寸（约22厘米）和声速（约340米/秒）计算，要实现3度角的辨别精度，需要约30微秒的时差判断。由于这种时差判断依赖于声波群反射原理，因此方向感知功能主要应用于频率低于1千赫兹的声音。

这两种方向性信息的来源都极大地受益于头部转动以观察信号变化的能力。当听者通过耳机接收完全相同的声音时（例如单耳听音），会产生一种有趣的感觉。大脑会判定该声音源自听者头部的中心位置！

人类的听觉虽然能判断声音的*来源方向*，但在识别*距离*方面却力不从心。这是因为声波中能提供这类信息的线索实在有限。人类听觉对高频声音感知较弱，而对低频声音则能感知其远近。这是因为声波在传播过程中会逐渐衰减高频成分。回声内容则是另一个较弱的距离线索，它能帮助我们感知房间的大小。例如，

sounds in a large auditorium will contain echoes at about 100 millisecond intervals, while 10 milliseconds is typical for a small office. Some species have solved this ranging problem by using *active sonar*. For example, bats and dolphins produce clicks and squeaks that reflect from nearby objects. By measuring the interval between transmission and echo, these animals can locate objects with about 1 cm resolution. Experiments have shown that some humans, particularly the blind, can also use active echo localization to a small extent.

Timbre

The perception of a continuous sound, such as a note from a musical instrument, is often divided into three parts: **loudness**, **pitch**, and **timbre** (pronounced "timber"). *Loudness* is a measure of sound wave intensity, as previously described. *Pitch* is the frequency of the fundamental component in the sound, that is, the frequency with which the waveform repeats itself. While there are subtle effects in both these perceptions, they are a straightforward match with easily characterized physical quantities.

Timbre is more complicated, being determined by the *harmonic content* of the signal. Figure 22-2 illustrates two waveforms, each formed by adding a 1 kHz sine wave with an amplitude of *one*, to a 3 kHz sine wave with an amplitude of *one-half*. The difference between the two waveforms is that the one shown in (b) has the higher frequency *inverted* before the addition. Put another way, the third harmonic (3 kHz) is phase shifted by 180 degrees compared to the first harmonic (1 kHz). In spite of the very different time domain waveforms, these two signals sound *identical*. This is because hearing is based on the *amplitude* of the frequencies, and is very insensitive to their *phase*. The *shape* of the time domain waveform is only indirectly related to hearing, and usually not considered in audio systems.

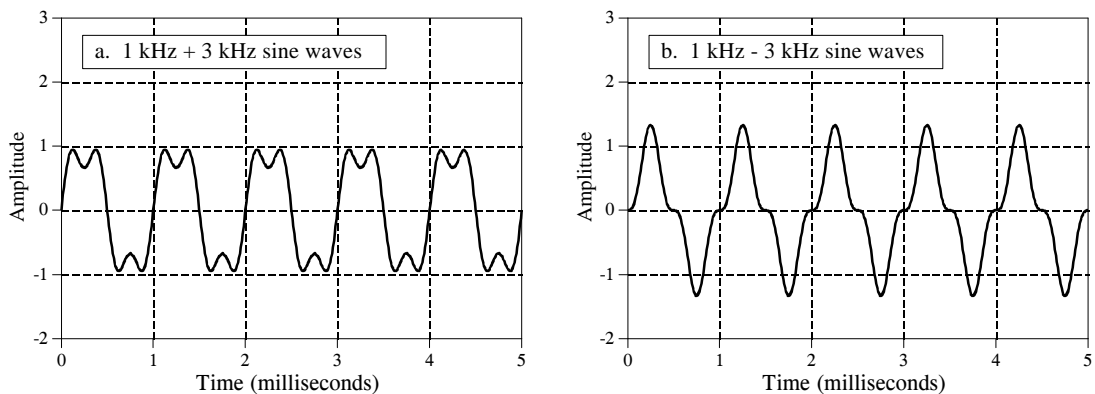


FIGURE 22-2

Phase detection of the human ear. The human ear is very insensitive to the relative phase of the component sinusoids. For example, these two waveforms would sound identical, because the *amplitudes* of their components are the same, even though their relative *phases* are different.

大型礼堂内的声音会产生间隔约100毫秒的回声，而小型办公室的回声间隔通常为10毫秒。部分物种通过使用*主动声呐*解决了这一测距难题。例如蝙蝠和海豚会发出咔嗒声和吱吱声，这些声音会在附近物体表面反射。通过测量声波发射与回声之间的间隔时间，这些动物能以约1厘米的精度定位物体。实验表明，部分人类（尤其是盲人）也能在一定程度上运用主动回声定位技术。

音色

对连续声音（如乐器音符）的感知通常被划分为三个部分：**响度**、**音高**和**音色**（发音为“timber”）。*响度*是声波强度的度量，如前所述。*音高*则是声音中基频成分的频率，即波形重复出现的频率。尽管这两种感知存在细微差异，但它们与易于表征的物理量有着直接对应关系。

*音色*更为复杂，由信号的*谐波成分*决定。图22-2展示了两种波形，每种波形都是将一个振幅为1的1 kHz正弦波叠加到一个振幅为1/2的3 kHz正弦波上。两种波形的区别在于所示波形在(b)中，较高频率的*倒置*会在加法运算前发生。换言之，第三谐波（3kHz）与第一谐波（1kHz）相比相位相差180度。尽管时域波形差异显著，这两种信号听起来却完全一致。这是因为人类听觉主要依赖频率的*幅度*，对*相位*敏感度极低。时域波形的*形状*与听觉感知仅存在间接关联，因此在音频系统设计中通常不予考虑。

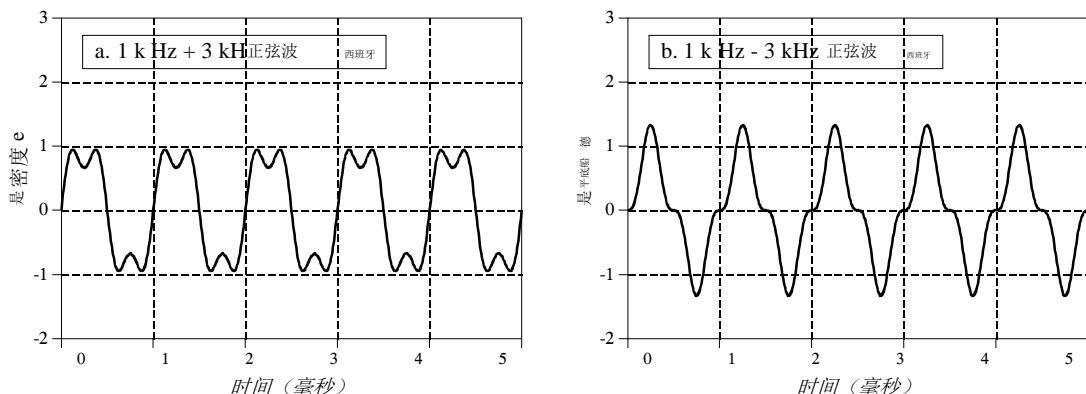


图22-2

人耳的相位检测。人耳对分量正弦波的相对相位非常不敏感。例如，这两个波形听起来是相同的，因为它们的分量*幅度*相同，尽管它们的相对*相位*不同。

The ear's insensitivity to phase can be understood by examining how sound propagates through the environment. Suppose you are listening to a person speaking across a small room. Much of the sound reaching your ears is reflected from the walls, ceiling and floor. Since sound propagation depends on frequency (such as: attenuation, reflection, and resonance), different frequencies will reach your ear through different paths. This means that the relative phase of each frequency will change as you move about the room. Since the ear disregards these phase variations, you perceive the voice as *unchanging* as you move position. From a physics standpoint, the phase of an audio signal becomes randomized as it propagates through a complex environment. Put another way, the ear is insensitive to phase because it contains little useful information.

However, it cannot be said that the ear is completely deaf to the phase. This is because a phase change can rearrange the *time sequence* of an audio signal. An example is the chirp system (Chapter 11) that changes an impulse into a much longer duration signal. Although they differ only in their phase, the ear can distinguish between the two sounds because of their difference in duration. For the most part, this is just a curiosity, not something that happens in the normal listening environment.

Suppose that we ask a violinist to play a note, say, the A below middle C. When the waveform is displayed on an oscilloscope, it appears much as the sawtooth shown in Fig. 22-3a. This is a result of the sticky rosin applied to the fibers of the violinist's bow. As the bow is drawn across the string, the waveform is formed as the string sticks to the bow, is pulled back, and eventually breaks free. This cycle repeats itself over and over resulting in the sawtooth waveform.

Figure 22-3b shows how this sound is perceived by the ear, a frequency of 220 hertz, plus harmonics at 440, 660, 880 hertz, etc. If this note were played on another instrument, the waveform would *look* different; however, the ear would still hear a frequency of 220 hertz plus the harmonics. Since the two instruments produce the same fundamental frequency for this note, they sound similar, and are said to have identical *pitch*. Since the relative amplitude of the *harmonics* is different, they will not sound identical, and will be said to have different *timbre*.

It is often said that timbre is determined by the shape of the waveform. This is true, but slightly misleading. The perception of timbre results from the ear detecting harmonics. While harmonic content is determined by the shape of the waveform, the insensitivity of the ear to phase makes the relationship very one-sided. That is, a particular waveform will have only one timbre, while a particular timbre has an infinite number of possible waveforms.

The ear is very accustomed to hearing a fundamental plus harmonics. If a listener is presented with the combination of a 1 kHz and 3 kHz sine wave, they will report that it sounds natural and pleasant. If sine waves of 1 kHz and 3.1 kHz are used, it will sound objectionable.

要理解耳朵对相位的不敏感性，可以通过观察声音在环境中的传播特性来说明。假设你在小房间内聆听他人说话，大部分传入耳中的声音都会被墙壁、天花板和地板反射。由于声波传播依赖频率（例如衰减、反射和共振），不同频率的声波会通过不同路径到达耳朵。这意味着当你在房间内移动时，每个频率的相对相位会发生变化。由于耳朵对这些相位变化视而不见，因此你会觉得声音在移动时 *保持不变*。从物理学角度看，音频信号在复杂环境中传播时，其相位会变得随机化。换句话说，耳朵之所以对相位不敏感，是因为相位信息本身包含的有用信息太少。

不过，我们不能说人耳对相位完全无感。这是因为相位变化会改变音频信号的 *时间序列*。以啁啾系统（第11章）为例，它能将短促的脉冲信号转化为持续时间显著延长的信号。虽然两者仅在相位上存在差异，但人耳能通过时长的不同来区分这两种声音。不过这主要是个有趣现象，并不会在日常听觉环境中实际发生。

假设我们请小提琴手演奏一个音符，比如中央音C下方的A。当波形显示在示波器上时，其形态与图22-3a所示的锯齿波极为相似。这是由于小提琴弓的纤维上涂有粘性松香所致。当弓弦被拉过琴弦时，波形的形成过程是：琴弦先粘附在弓上，随后被拉回，最终挣脱。这个循环不断重复，最终形成了锯齿波形。

图22-3b展示了人耳对220赫兹基频及其440、660、880赫兹谐波的感知效果。若将同一音符用不同乐器演奏，波形会 *显现* 差异，但人耳仍能分辨出220赫兹基频与谐波成分。由于两种乐器产生的基频完全一致，因此音色相似，被称作具有相同 *音高*。而 *谐波* 的相对振幅存在差异，导致音色产生明显区别，这种差异被称为不同的 *音色*。

人们常说音色由波形决定，这虽正确却存在偏差。实际上，人耳感知音色源于对谐波的识别。虽然谐波成分由波形决定，但因人耳对相位不敏感，这种关联呈现单向性——即特定波形对应单一音色，而特定音色则对应无限可能的波形。

人耳对基频加谐波的听觉非常适应。当听众听到1 kHz与3 kHz正弦波的组合时，会反馈其听觉效果自然悦耳；若使用1 kHz与3.1 kHz正弦波，则会感到不适。

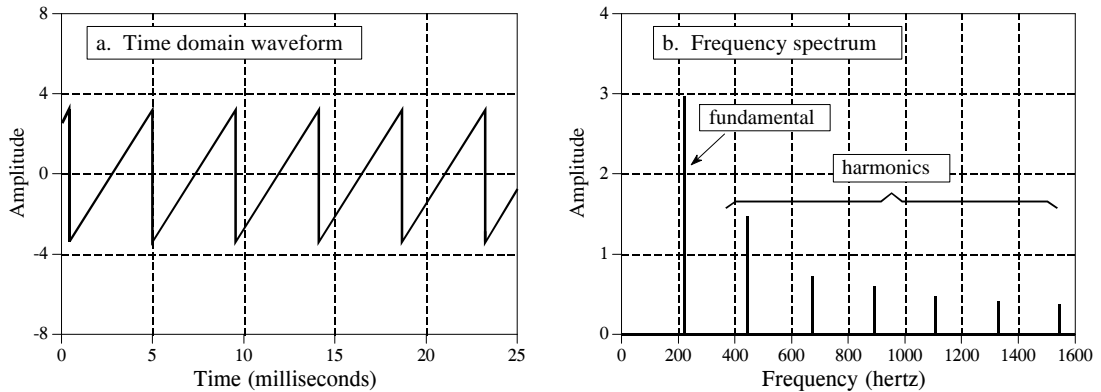


FIGURE 22-3

Violin waveform. A bowed violin produces a sawtooth waveform, as illustrated in (a). The sound heard by the ear is shown in (b), the fundamental frequency plus harmonics.

This is the basis of the standard musical scale, as illustrated by the piano keyboard in Fig. 22-4. Striking the farthest left key on the piano produces a fundamental frequency of 27.5 hertz, plus harmonics at 55, 110, 220, 440, 880 hertz, etc. (there are also harmonics between these frequencies, but they aren't important for this discussion). These harmonics correspond to the fundamental frequency produced by other keys on the keyboard. Specifically, every *seventh* white key is a harmonic of the far left key. That is, the eighth key from the left has a fundamental frequency of 55 hertz, the 15th key has a fundamental frequency of 110 hertz, etc. Being harmonics of each other, these keys sound similar when played, and are harmonious when played in unison. For this reason, they are *all* called the note, A. In this same manner, the white key immediate right of each A is called a B, and *they* are all harmonics of each other. This pattern repeats for the seven notes: A, B, C, D, E, F, and G.

The term **octave** means a *factor of two in frequency*. On the piano, one octave comprises eight white keys, accounting for the name (*octo* is Latin for *eight*). In other words, the piano's frequency doubles after every seven white keys, and the entire keyboard spans a little over seven octaves. The range of human hearing is generally quoted as 20 hertz to 20 kHz,

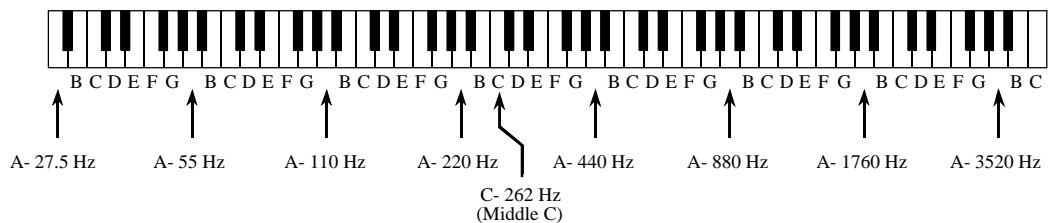


FIGURE 22-4

The Piano keyboard. The keyboard of the piano is a *logarithmic* frequency scale, with the fundamental frequency doubling after every seven white keys. These white keys are the notes: A, B, C, D, E, F and G.

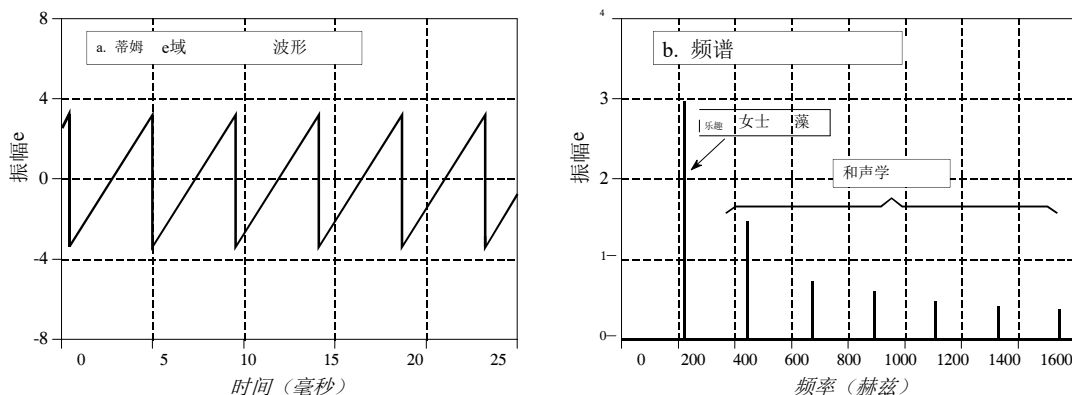


图22-3

小提琴波形。如(a)所示，拉弓小提琴产生锯齿波形。耳人听到的声音如(b)所示，即基频加谐波。

这就是标准音阶的基础，如图22-4中的钢琴键盘所示。按下钢琴最左边的键会产生27.5赫兹的基频，以及55、110、220、440、880赫兹等次谐波（这些频率之间还有其他谐波，但本讨论不涉及）。这些谐波对应着键盘上其他键产生的基频。具体来说，每个白键的第七个键都是最左边键的谐波。也就是说，从左数第八个键的基频是55赫兹，第十五个键的基频是110赫兹，以此类推。由于这些键彼此是谐波，演奏时声音相似，齐奏时也和谐统一。因此，它们都被称为音符A。同理，每个A键右侧的白键称为B键，它们之间也都是谐波关系。这个模式在七个音符中重复：A、B、C、D、E、F和G。

八度指的是频率的两倍倍数。在钢琴上，一个八度包含八个白键，因此得名（octo是拉丁语八的缩写）。换言之，钢琴的频率每经过七个白键就会翻倍，整个键盘的音域略超过七个八度。人类听觉范围通常被定义为20赫兹到20千赫兹。

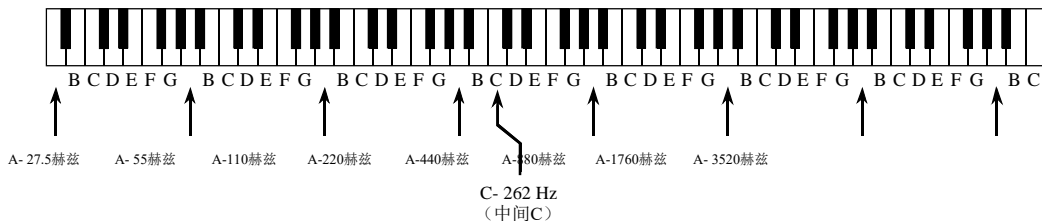


图22-4

钢琴键盘。钢琴键盘是一个对数频率标度，基频在每七个白键后倍增。这些白键是音符：A、B、C、D、E、F和G。

corresponding to about $\frac{1}{2}$ octave to the left, and two octaves to the right of the piano keyboard. Since octaves are based on doubling the frequency every fixed number of keys, they are a *logarithmic* representation of frequency. This is important because audio information is generally distributed in this same way. For example, as much audio information is carried in the octave between 50 hertz and 100 hertz, as in the octave between 10 kHz and 20 kHz. Even though the piano only covers about 20% of the frequencies that humans can hear (4 kHz out of 20 kHz), it can produce more than 70% of the audio information that humans can perceive (7 out of 10 octaves). Likewise, the highest frequency a human can detect drops from about 20 kHz to 10 kHz over the course of an adult's lifetime. However, this is only a loss of about 10% of the hearing ability (one octave out of ten). As shown next, this logarithmic distribution of information directly affects the required *sampling rate* of audio signals.

Sound Quality vs. Data Rate

When designing a digital audio system there are two questions that need to be asked: (1) how good does it need to sound? and (2) what data rate can be tolerated? The answer to these questions usually results in one of three categories. First, **high fidelity music**, where sound quality is of the greatest importance, and almost any data rate will be acceptable. Second, **telephone communication**, requiring natural sounding speech *and* a low data rate to reduce the system cost. Third, **compressed speech**, where reducing the data rate is very important and some unnaturalness in the sound quality can be tolerated. This includes military communication, cellular telephones, and digitally stored speech for voice mail and multimedia.

Table 22-2 shows the tradeoff between sound quality and data rate for these three categories. High fidelity music systems sample fast enough (44.1 kHz), and with enough precision (16 bits), that they can capture virtually all of the sounds that humans are capable of hearing. This magnificent sound quality comes at the price of a high data rate, $44.1 \text{ kHz} \times 16 \text{ bits} = 706 \text{ k bits/sec}$. This is pure brute force.

Whereas music requires a bandwidth of 20 kHz, natural sounding speech only requires about 3.2 kHz. Even though the frequency range has been reduced to only 16% (3.2 kHz out of 20 kHz), the signal still contains 80% of the original sound information (8 out of 10 octaves). Telecommunication systems typically operate with a sampling rate of about 8 kHz, allowing natural sounding speech, but greatly reduced music quality. You are probably already familiar with this difference in sound quality: FM radio stations broadcast with a bandwidth of almost 20 kHz, while AM radio stations are limited to about 3.2 kHz. Voices sound normal on the AM stations, but the music is weak and unsatisfying.

Voice-only systems also reduce the precision from 16 bits to 12 bits per sample, with little noticeable change in the sound quality. This can be reduced to only 8 bits per sample if the quantization step size is made unequal. This is a widespread procedure called **companding**, and will be

在钢琴键盘左侧约 $\frac{1}{2}$ 个八度音程，右侧则有两个八度音程。由于八度音程的频率倍增基于固定键数的倍增规律，因此它们是频率的 *对数* 表示法。这一点至关重要，因为音频信息通常以相同方式分布。例如，50赫兹至100赫兹之间的八度音程所承载的音频信息量，与10千赫至20千赫之间的八度音程相当。尽管钢琴仅覆盖人类可听频率的约20%（20千赫中的4千赫），却能产生人类可感知音频信息的70%以上（10个八度中的7个）。同样地，成年人一生中可检测的最高频率会从约20千赫降至10千赫。但这仅相当于听力能力损失约10%（十个八度中的一个）。如下所示，这种信息的对数分布直接影响音频信号所需的采样率。

音质与数据速率

在设计数字音频系统时，需要考虑两个核心问题：(1)系统需要达到怎样的音质水平？(2)能接受多高的数据传输速率？通常答案会归为以下三类：第一类是**高保真音乐**，这类场景对音质要求极高，几乎任何数据传输速率都能满足需求；第二类是**电话通信**，需要保持语音自然、且采用低数据速率以降低系统成本；第三类是**压缩语音**，此时降低数据速率至关重要，但可以容忍一定程度的音质失真。这类应用场景包括军事通信、移动电话，以及用于语音信箱和多媒体的数字存储语音。

表22-2展示了这三类系统在音质与数据速率间的权衡。高保真音乐系统采样频率高达44.1kHz，精度达16位，几乎能捕捉人类可感知的所有声音。这种卓越音质的代价是极高的数据速率—— $44.1\text{kHz} \times 16\text{位} = 706\text{k bits/sec}$ 。这简直是纯粹的硬核配置。

音乐需要20千赫的带宽，而自然人声仅需约3.2千赫。虽然频率范围已缩减至16%（20千赫中的3.2千赫），但信号仍保留了80%的原始声音信息（相当于10个八度中的8个）。电信系统通常采用约8千赫的采样率，虽然能播放自然人声，但音乐音质会大幅降低。您可能已经注意到这种音质差异：调频电台的广播带宽接近20千赫，而调幅电台则受限于约3.2千赫。人声在调幅电台听起来正常，但音乐音量微弱且不够悦耳。

语音系统会将采样精度从16位降低到12位，但音质变化不大。若采用不等量化步长，精度甚至可降至8位。这种技术称为**压缩扩频**，目前应用广泛。

Sound Quality Required	Bandwidth	Sampling rate	Number of bits	Data rate (bits/sec)	Comments
High fidelity music (compact disc)	5 Hz to 20 kHz	44.1 kHz	16 bit	706k	Satisfies even the most picky audiophile. Better than human hearing.
Telephone quality speech (with companding)	200 Hz to 3.2 kHz	8 kHz	12 bit	96k	Good speech quality, but very poor for music.
	200 Hz to 3.2 kHz	8 kHz	8 bit	64k	Nonlinear ADC reduces the data rate by 50%. A very common technique.
Speech encoded by Linear Predictive Coding	200 Hz to 3.2 kHz	8 kHz	12 bit	4k	DSP speech compression technique. Very low data rates, poor voice quality.

TABLE 22-2

Audio data rate vs. sound quality. The sound quality of a digitized audio signal depends on its *data rate*, the product of its sampling rate and number of bits per sample. This can be broken into three categories, high fidelity music (706 kbits/sec), telephone quality speech (64 kbits/sec), and compressed speech (4 kbits/sec).

discussed later in this chapter. An 8 kHz sampling rate, with an ADC precision of 8 bits per sample, results in a data rate of 64k bits/sec. This is the *brute force* data rate for natural sounding speech. Notice that speech requires less than 10% of the data rate of high fidelity music.

The data rate of 64k bits/sec represents the straightforward application of sampling and quantization theory to audio signals. Techniques for lowering the data rate further are based on *compressing* the data stream by removing the inherent redundancies in speech signals. Data compression is the topic of Chapter 27. One of the most efficient ways of compressing an audio signal is **Linear Predictive Coding (LPC)**, of which there are several variations and subgroups. Depending on the speech quality required, LPC can reduce the data rate to as little as 2-6k bits/sec. We will revisit LPC later in this chapter with *speech synthesis*.

High Fidelity Audio

Audiophiles demand the utmost sound quality, and all other factors are treated as secondary. If you had to describe the mindset in one word, it would be: *overkill*. Rather than just matching the abilities of the human ear, these systems are designed to *exceed* the limits of hearing. It's the only way to be sure that the reproduced music is pristine. Digital audio was brought to the world by the **compact laser disc**, or **CD**. This was a revolution in music; the sound quality of the CD system far exceeds older systems, such as records and tapes. DSP has been at the forefront of this technology.

所需音质	带宽	抽样 比率	数量 位数	数据速率 比特/秒	评论
高保真音乐（CD）	5 赫兹至 20千赫	44.1千赫	16位	706k	连最挑剔的发烧友也能满足， 其表现甚至超越人类听觉。
电话语音质量 （含压缩）	200 赫兹至 3.2千赫	8千赫	12位	96k	语音质量良好，但音乐播放 效果极差。
	200 赫兹至 3.2千赫	8千赫	8位	64k	非线性模数转换器（ADC）可 将数据速率降低50%，是一种 非常常用的技术。
线性预测编码语音	200 赫兹至 3.2千赫	8千赫	12位	4k	数字语音编码技术。数据传 输速率极低，语音质量较 差。

表22-2
音频数据速率与音质。数字化音频信号的音质取决于其数据速率，即采样率与每样本比特数的乘积。这可分为三类：高保真音乐（706 kbits/sec）、电话质量语音（64 kbits/sec）和压缩语音（4 kbits/sec）。

8kHz的采样率，每个样本的ADC精度为8位，数据速率为64kbit/s。这是自然语音的原始数据速率。注意，语音所需的数据速率不到高保真音乐的10%。

64千比特/秒的数据传输速率，是将采样与量化理论直接应用于音频信号的典型成果。要进一步降低数据速率，关键在于通过去除语音信号中的固有冗余信息来实现数据压缩。数据压缩技术将在第27章详细探讨，其中线性预测编码（LPC）作为最高效的音频压缩方案之一，已发展出多种改进版本和子类。根据语音质量要求的不同，LPC技术可将数据速率压缩至2-6千比特/秒的低水平。本章后续章节将结合语音合成技术，对LPC技术进行深入解析。

高保真音频

音响发烧友追求极致音质，其他因素都退居次席。若要用一个词形容这种心态，那就是：过度追求。这些系统不仅满足人耳的听觉需求，更致力于超越听觉极限。唯有如此，才能确保音乐还原得完美无瑕。数字音频技术的诞生要归功于光盘（即CD），这堪称音乐界的革命性突破。CD系统的音质远超黑胶唱片、磁带等传统介质。数字信号处理技术始终引领着这一技术浪潮。

Figure 22-5 illustrates the surface of a compact laser disc, such as viewed through a high power microscope. The main surface is shiny (reflective of light), with the digital information stored as a series of dark pits burned on the surface with a laser. The information is arranged in a single track that spirals from the outside to the inside, the same as a phonograph record. The rotation of the CD is changed from about 210 to 480 rpm as the information is read from the outside to the inside of the spiral, making the scanning velocity a constant 1.2 meters per second. (In comparison, phonograph records spin at a fixed rate, such as 33, 45 or 78 rpm). During playback, an optical sensor detects if the surface is reflective or nonreflective, generating the corresponding binary information.

As shown by the geometry in Fig. 22-5, the CD stores about 1 bit per $(\mu\text{m})^2$, corresponding to 1 million bits per $(\text{mm})^2$, and 15 billion bits per disk. This is about the same feature size used in integrated circuit manufacturing, and for a good reason. One of the properties of light is that it cannot be focused to smaller than about one-half wavelength, or $0.3 \mu\text{m}$. Since both integrated circuits and laser disks are created by optical means, the fuzziness of light below $0.3 \mu\text{m}$ limits how small of features can be used.

Figure 22-6 shows a block diagram of a typical compact disc playback system. The raw data rate is 4.3 million bits per second, corresponding to 1 bit each $0.28 \mu\text{m}$ of track length. However, this is in conflict with the specified geometry of the CD; each pit must be no shorter than $0.8 \mu\text{m}$, and no longer than $3.5 \mu\text{m}$. In other words, each binary *one* must be part of a group of 3 to 13 *ones*. This has the advantage of reducing the error rate due to the optical pickup, but how do you force the binary data to comply with this strange bunching?

The answer is an encoding scheme called **eight-to-fourteen modulation (EFM)**. Instead of directly storing a byte of data on the disc, the 8 bits are passed through a look-up table that pops out 14 bits. These 14 bits have the desired bunching characteristics, and are stored on the laser disc. Upon playback, the binary values read from the disc are passed through the inverse of the EFM look-up table, resulting in each 14 bit group being turned back into the correct 8 bits.

FIGURE 22-5

Compact disc surface. Micron size pits are burned into the surface of the CD to represent ones and zeros. This results in a data density of 1 bit per μm^2 , or one million bits per mm^2 . The pit depth is $0.16 \mu\text{m}$.

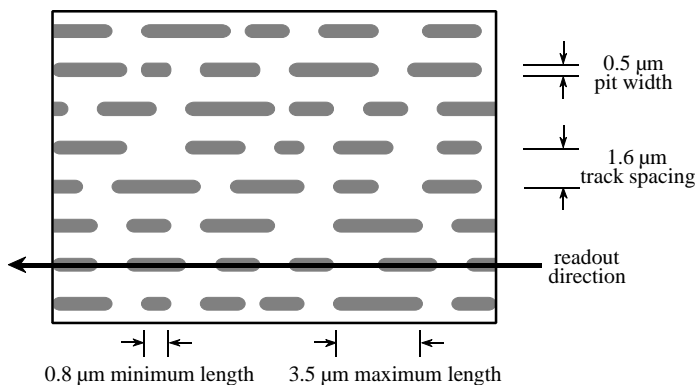


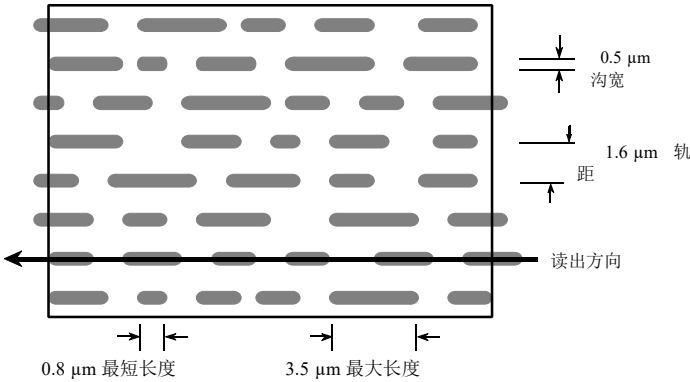
图22-5展示了通过高倍显微镜观察的紧凑型激光光盘表面结构。主表面具有高光泽度（能反射光线），其数字信息以激光烧蚀形成的深色凹坑形式存储。这些信息沿螺旋轨道排列，从外向内螺旋延伸，与黑胶唱片的结构相同。当从螺旋轨道外侧读取信息时，光盘转速会从约210转/分钟调整至480转/分钟，使扫描速度保持恒定的1.2米/秒。（相比之下，黑胶唱片转速固定为33、45或78转/分钟）。在播放过程中，光学传感器会检测表面是否具有反射特性，并生成对应的二进制信息。

如图22-5的几何结构所示，光盘存储密度约为每平方（ μm ）² 1比特²，相当于每平方毫米（毫米）² 存储100万比特，单个光盘可存储150亿比特。这一特征尺寸与集成电路制造中使用的尺寸相当，且有其合理依据。光的特性之一是其聚焦范围无法小于约半波长（即 $0.3\ \mu\text{m}$ ）。由于集成电路和激光光盘均通过光学手段制造， $0.3\ \mu\text{m}$ 以下的光的模糊特性限制了可采用的特征尺寸。

图22-6展示了典型光盘播放系统的框图。原始数据速率为每秒430万比特，相当于每 $0.28\ \mu\text{m}$ 的轨道长度对应1比特。但这一参数与光盘的几何规格存在冲突——每个凹坑的深度必须在 $0.8\ \mu\text{m}$ 至 $3.5\ \mu\text{m}$ 之间。换句话说，每个二进制1必须包含在由3到13个1组成的簇中。这种设计虽然能有效降低光学拾取头的误读率，但如何强制二进制数据遵循这种特殊的簇化规则呢？

答案是一种名为**八到十四调制（EFM）**的编码方案。该方案并非直接将一个字节数据存储光盘上，而是将8位数据通过查找表转换为14位数据。这14位数据具有所需的分组特性，并存储在激光光盘上。在播放时，从光盘读取的二进制值通过 EFM 查找表的逆向转换，使每个14位数据组重新还原为正确的8位数据。

图22-5
光盘表面。微米级凹坑被刻录在光盘表面以表示1和0。这使得数据密度达到每平方 μm 1比特²，即每平方毫米一百万比特²。凹坑深度为 $0.16\ \mu\text{m}$ 。



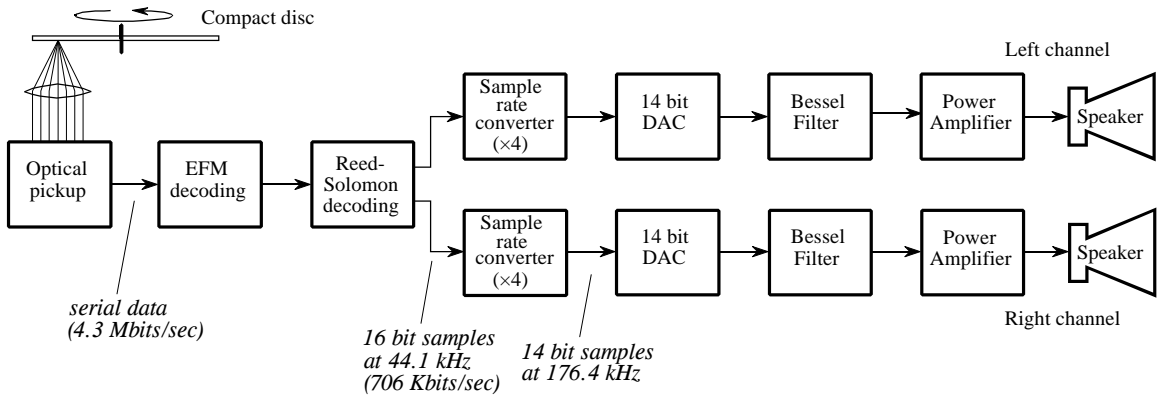


FIGURE 22-6

Compact disc playback block diagram. The digital information is retrieved from the disc with an optical sensor, corrected for EFM and Reed-Solomon encoding, and converted to stereo analog signals.

In addition to EFM, the data are encoded in a format called **two-level Reed-Solomon coding**. This involves combining the left and right stereo channels along with data for error detection and correction. Digital errors detected during playback are either: *corrected* by using the redundant data in the encoding scheme, *concealed* by interpolating between adjacent samples, or *muted* by setting the sample value to zero. These encoding schemes result in the data rate being *tripled*, i.e., 1.4 Mbits/sec for the stereo audio signals versus 4.3 Mbits/sec stored on the disc.

After decoding and error correction, the audio signals are represented as 16 bit samples at a 44.1 kHz sampling rate. In the simplest system, these signals could be run through a 16 bit DAC, followed by a low-pass analog filter. However, this would require high performance analog electronics to pass frequencies below 20 kHz, while rejecting all frequencies above 22.05 kHz, $\frac{1}{2}$ of the sampling rate. A more common method is to use a **multirate** technique, that is, convert the digital data to a higher sampling rate before the DAC. A factor of four is commonly used, converting from 44.1 kHz to 176.4 kHz. This is called **interpolation**, and can be explained as a two step process (although it may not actually be carried out this way). First, three samples with a value of zero are placed between the original samples, producing the higher sampling rate. In the frequency domain, this has the effect of duplicating the 0 to 22.05 kHz spectrum three times, at 22.05 to 44.1 kHz, 44.1 to 66.15 kHz, and 66.15 to 88.2 kHz. In the second step, an efficient *digital* filter is used to remove the newly added frequencies.

The sample rate increase makes the sampling interval smaller, resulting in a smoother signal being generated by the DAC. The signal still contains frequencies between 20 Hz and 20 kHz; however, the Nyquist frequency has been increased by a factor of four. This means that the analog filter only needs to pass frequencies below 20 kHz, while blocking frequencies above 88.2 kHz. This is usually done with a three pole Bessel filter. Why use a *Bessel* filter if the ear is insensitive to phase? Overkill, remember?

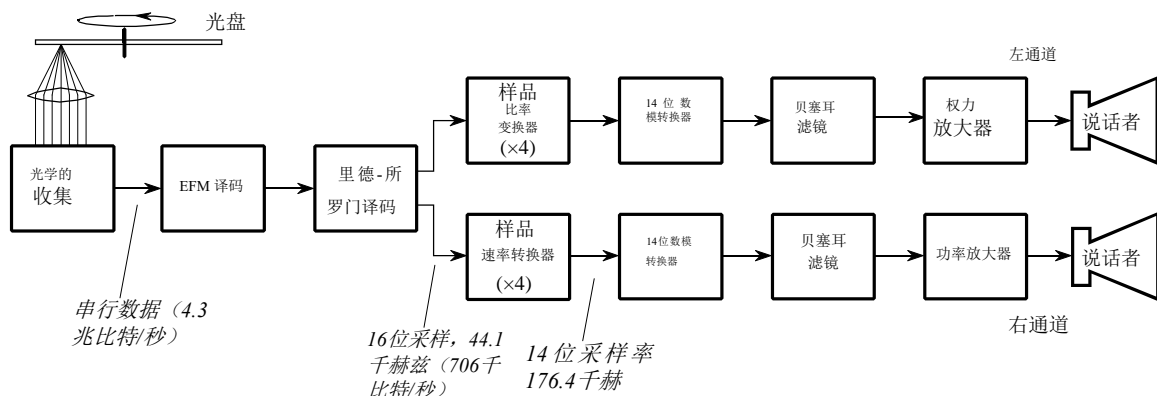


图22-6

光盘播放系统框图。数字信息通过光学传感器从光盘读取，经 EFM 校正和里德-所罗门编码处理后，转换为立体声模拟信号。

除了 EFM 之外，这些数据还采用了名为**双层里德-所罗门编码**的格式。该编码方案将左右声道数据与错误检测校正信息相结合。播放过程中检测到的数字错误可通过三种方式处理：*利用编码方案中的冗余数据进行校正、通过相邻样本插值进行隐藏，或将样本值设为零进行静音处理*。这种编码方案使数据传输速率提升三倍——立体声信号的传输速率从4.3兆比特/秒提升至1.4兆比特/秒，而光盘存储的原始数据速率则为4.3兆比特/秒。

经过解码和纠错处理后，音频信号以16位采样、44.1kHz的采样率呈现。在最简单的系统中，这些信号可先通过16位数模转换器（DAC），再经过低通模拟滤波器处理。但这种方法需要高性能模拟电子元件才能通过20 kHz以下的频率，同时完全阻隔高于采样率 $\frac{1}{2}$ 22.05kHz的信号。更常见的做法是采用**多速率**技术，即在数模转换前将数字数据转换为更高采样率。通常采用四倍率转换，将44.1kHz提升至176.4kHz。这种技术称为**插值**，其原理可分解为两个步骤（尽管实际操作未必完全如此）。首先，在原始采样点之间插入三个零值采样点，从而获得更高采样率。在频域中，这相当于将0-22.05kHz频谱重复三次，分别对应22.05-44.1kHz、41-66.15kHz和66.15-88.2kHz。第二步则通过高效数字滤波器去除新增的高频成分。

提高采样率会缩短采样间隔，使得数模转换器生成的信号更加平滑。虽然信号仍包含20赫兹到20千赫兹之间的频率，但奈奎斯特频率已提升至原来的四倍。这意味着模拟滤波器只需通过20千赫兹以下的频率，同时阻断88.2千赫兹以上的频率。这通常通过三极贝塞尔滤波器来实现。既然人耳对相位不敏感，为何还要使用贝塞尔滤波器？这波操作是不是有点过头了？

Since there are four times as many samples, the number of bits per sample can be reduced from 16 bits to 14 bits, without degrading the sound quality. The $\sin(x)/x$ correction needed to compensate for the zeroth order hold of the DAC can be part of either the analog or digital filter.

Audio systems with more than one channel are said to be in **stereo** (from the Greek word for *solid*, or *three-dimensional*). Multiple channels send sound to the listener from different directions, providing a more accurate reproduction of the original music. Music played through a monaural (one channel) system often sounds artificial and bland. In comparison, a good stereo reproduction makes the listener feel as if the musicians are only a few feet away. Since the 1960s, high fidelity music has used two channels (left and right), while motion pictures have used four channels (left, right, center, and surround). In early stereo recordings (say, the Beatles or the Mamas And The Papas), individual singers can often be heard in only one channel or the other. This rapidly progressed into a more sophisticated **mix-down**, where the sound from many microphones in the recording studio is combined into the two channels. Mix-down is an art, aimed at providing the listener with the perception of *being there*.

The four channel sound used in motion pictures is called **Dolby Stereo**, with the home version called **Dolby Surround Pro Logic**. ("Dolby" and "Pro Logic" are trademarks of Dolby Laboratories Licensing Corp.). The four channels are encoded into the standard left and right channels, allowing regular two-channel stereo systems to reproduce the music. A Dolby decoder is used during playback to recreate the four channels of sound. The left and right channels, from speakers placed on each side of the movie or television screen, is similar to that of a regular two-channel stereo system. The speaker for the center channel is usually placed directly above or below the screen. Its purpose is to reproduce speech and other visually connected sounds, keeping them firmly centered on the screen, regardless of the seating position of the viewer/listener. The surround speakers are placed to the left and right of the listener, and may involve as many as twenty speakers in a large auditorium. The surround channel only contains midrange frequencies (say, 100 Hz to 7 kHz), and is *delayed* by 15 to 30 milliseconds. This delay makes the listener perceive that speech is coming from the screen, and not the sides. That is, the listener hears the speech coming from the front, followed by a delayed version of the speech coming from the sides. The listener's mind interprets the delayed signal as a reflection from the walls, and ignores it.

Companding

The data rate is important in telecommunication because it is directly proportional to the *cost* of transmitting the signal. Saving bits is the same as saving money. **Companding** is a common technique for reducing the data rate of audio signals by making the quantization levels *unequal*. As previously mentioned, the loudest sound that can be tolerated (120 dB SPL) is about one-million times the amplitude of the weakest sound that can be detected (0 dB

由于样本数量是原来的四倍，每个样本的比特数可以从16位减少到14位，而不会降低音质。补偿DAC零阶保持所需的 $\sin(x)/x$ 校正可以是模拟滤波器或数字滤波器的一部分。

多声道音频系统被称为**立体声**（源自希腊语 $solid$ ，意为三维）。多声道系统通过不同方向的声源，能更精准还原音乐原声。单声道（单声道）播放的音乐常显得生硬平淡，而优质的立体声效果能让听众仿佛置身音乐家身旁。自20世纪60年代起，高保真音乐采用左右声道，电影则使用四声道（左、右、中、环绕）。早期立体声录音（如披头士或妈妈与爸爸乐队的作品）中，歌手的声线常仅能通过单声道呈现。这种技术很快发展为更复杂的**混音技术**，将录音棚内多个麦克风采集的声源整合为双声道。混音是一门艺术，旨在让听众产生身临其境的听觉体验。

电影中使用的四声道声音被称为**杜比立体声**，家用版本则称为**杜比环绕声专业逻辑**（“杜比”和“专业逻辑”均为杜比实验室授权公司注册商标）。这四个声道被编码为标准的左右声道，使得普通双声道立体声系统也能还原音乐效果。播放时需使用杜比解码器来重建四声道声音。左右声道通过放置在电影或电视屏幕两侧的扬声器呈现，与普通双声道系统类似。中央声道的扬声器通常直接置于屏幕正上方或正下方，其作用是还原语音和其他视觉相关声音，确保无论观众/听众坐在屏幕的哪个位置，这些声音都能稳稳地居中呈现。环绕声扬声器则布置在听众左右两侧，在大型礼堂中可能多达二十个。环绕声道仅包含中频段（例如100赫兹至7千赫），并会进行15至30毫秒的**延迟处理**。这种延迟使听者感知到语音来自屏幕而非两侧。即听者先听到前方传来的语音，随后是来自两侧的延迟语音。听者大脑将延迟信号解释为墙壁的反射，并忽略其存在。

压缩-扩张

数据速率在电信领域至关重要，因为它与信号传输的成本成正比。节省比特数就等于节省资金。**压缩编码**是通过使量化级不等来降低音频信号数据速率的常用技术。如前所述，可容忍的最大声音（120分贝声压级）约为可检测到的最弱声音（0分贝）振幅的百万分之一。

SPL). However, the ear cannot distinguish between sounds that are closer than about 1 dB (12% in amplitude) apart. In other words, there are only about 120 different loudness levels that can be detected, spaced logarithmically over an amplitude range of one-million.

This is important for digitizing audio signals. If the quantization levels are equally spaced, 12 bits must be used to obtain telephone quality speech. However, only 8 bits are required if the quantization levels are made *unequal*, matching the characteristics of human hearing. This is quite intuitive: if the signal is small, the levels need to be very close together; if the signal is large, a larger spacing can be used.

Comping can be carried out in three ways: (1) run the analog signal through a nonlinear circuit before reaching a linear 8 bit ADC, (2) use an 8 bit ADC that internally has unequally spaced steps, or (3) use a linear 12 bit ADC followed by a digital look-up table (12 bits in, 8 bits out). Each of these three options requires the same nonlinearity, just in a different place: an analog circuit, an ADC, or a digital circuit.

Two nearly identical standards are used for companding curves: **μ255 law** (also called **mu law**), used in North America, and **"A" law**, used in Europe. Both use a logarithmic nonlinearity, since this is what converts the spacing detectable by the human ear into a linear spacing. In equation form, the curves used in μ255 law and "A" law are given by:

EQUATION 22-1

Mu law companding. This equation provides the nonlinearity for μ255 law companding. The constant, μ , has a value of 255, accounting for the name of this standard.

$$y = \frac{\ln(1 + \mu x)}{\ln(1 + \mu)} \quad \text{for } 0 \leq x \leq 1$$

EQUATION 22-2

"A" law companding. The constant, A , has a value of 87.6.

$$y = \frac{1 + \ln(Ax)}{1 + \ln(A)} \quad \text{for } 1/A \leq x \leq 1$$

$$y = \frac{Ax}{1 + \ln(A)} \quad \text{for } 0 \leq x \leq 1/A$$

Figure 22-7 graphs these equations for the input variable, x , being between -1 and +1, resulting in the output variable also assuming values between -1 and +1. Equations 22-1 and 22-2 only handle positive input values; portions of the curves for negative input values are found from symmetry. As shown in (a), the curves for μ255 law and "A" law are nearly identical. The only significant difference is near the origin, shown in (b), where μ255 law is a smooth curve, and "A" law switches to a straight line.

Producing a stable nonlinearity is a difficult task for analog electronics. One method is to use the logarithmic relationship between current and

然而，人耳无法分辨相距小于约1分贝（12%振幅）的声音。换言之，在振幅范围为一百万倍的情况下，仅能检测到约120个不同的响度水平，这些水平呈对数分布。

这对音频信号的数字化处理至关重要。若量化级数等距分布，需使用12位才能实现电话级语音质量。但若采用不等距量化，只需8位即可满足人耳听觉特性。这种设计逻辑非常直观：当信号较弱时，量化级数需紧密排列；当信号较强时，则可适当增大间隔。

压缩扩展技术可通过三种方式实现：(1) 使模拟信号在进入线性8位模数转换器前先经过非线性电路处理；(2) 使用内部具有非等距步进的8位模数转换器；(3) 采用线性12位模数转换器后接数字查找表（输入12位，输出8位）。这三种方案虽在实现位置上不同（模拟电路、模数转换器或数字电路），但均需相同的非线性处理。

压缩曲线使用两种几乎相同的标准：**μ255 定律**（也称**μ定律**），用于北美，以及**“A定律”**，用于欧洲。两者都使用对数非线性，因为这是将人耳可检测的间距转换为线性间距的方法。用方程式表示，μ255 定律和“A定律”使用的曲线由以下形式给出：

方程22-1

μ 律压扩。该方程为μ255 律压扩提供了非线性特性。常数 μ 的取值为 255，这也是该标准命名的由来。

$$y = \frac{\ln(1+\mu x)}{\ln(1+\mu)} \quad \text{for } 0 \leq x \leq 1$$

方程22-2

A 律压扩。常数 A 的取值为 87.6。

$$y = \frac{1+\ln(Ax)}{1+\ln(A)} \quad \text{for } 1/A \leq x \leq 1$$

$$y = \frac{Ax}{1+\ln(A)} \quad \text{for } 0 \leq x \leq 1/A$$

图22-7展示了输入变量x在-1到+1区间变化时对应的方程曲线，输出变量同样呈现-1到+1的取值范围。方程22-1和22-2仅处理正值输入，负值输入对应的曲线部分通过对称性推导得出。如(a)所示，μ255 定律与“A”定律的曲线几乎完全重合。唯一显著差异出现在原点附近区域（如(b)所示），该处 μ255 定律呈现平滑曲线，而“A”定律则转变为直线形态。

在模拟电子学中，产生稳定的非线性关系是一项艰巨的任务。一种方法是利用电流与电压之间的对数关系。

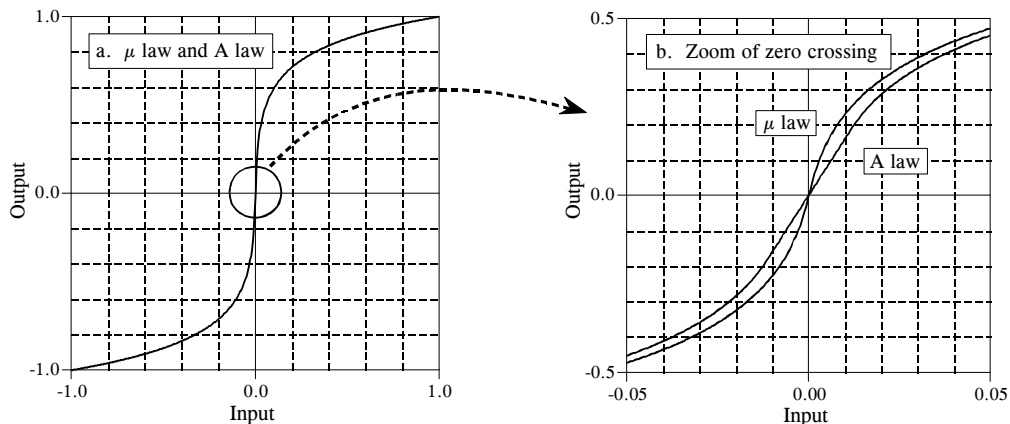


FIGURE 22-7

Companding curves. The μ 255 law and "A" law companding curves are nearly identical, differing only near the origin. Companding increases the amplitude when the signal is small, and decreases it when it is large.

voltage across a pn diode junction, and then add circuitry to correct for the ghastly temperature drift. Most companding circuits take another strategy: approximate the nonlinearity with a group of straight lines. A typical scheme is to approximate the logarithmic curve with a group of 16 straight segments, called **cords**. The first bit of the 8 bit output indicates if the input is positive or negative. The next three bits identify which of the 8 positive or 8 negative cords is used. The last four bits break each cord into 16 equally spaced increments. As with most integrated circuits, companding chips have sophisticated and proprietary internal designs. Rather than worrying about what goes on inside of the chip, pay the most attention to the pinout and the specification sheet.

Speech Synthesis and Recognition

Computer generation and recognition of speech are formidable problems; many approaches have been tried, with only mild success. This is an active area of DSP research, and will undoubtedly remain so for many years to come. You will be very disappointed if you are expecting this section to describe how to build speech synthesis and recognition circuits. Only a brief introduction to the typical approaches can be presented here. Before starting, it should be pointed out that most commercial products that produce human sounding speech do not *synthesize* it, but merely play back a digitally recorded segment from a human speaker. This approach has great sound quality, but it is limited to the prerecorded words and phrases.

Nearly all techniques for speech synthesis and recognition are based on the model of human speech production shown in Fig. 22-8. Most human speech sounds can be classified as either **voiced** or **fricative**. Voiced sounds occur when air is forced from the lungs, through the vocal cords, and out of the mouth and/or nose. The vocal cords are two thin flaps of tissue stretched across

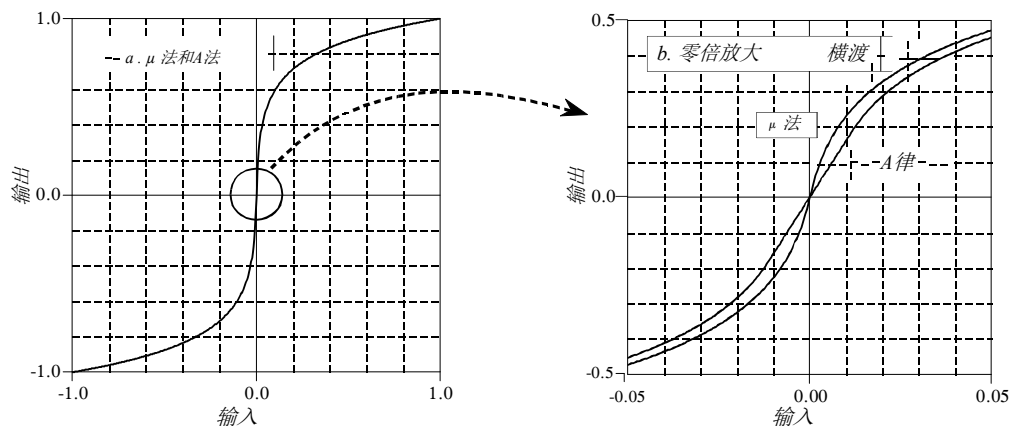


图22-7 压缩补偿曲线。 μ 255 定律与 ‘A’ 定律的压缩补偿曲线几乎完全相同，仅在原点附近存在差异。当信号较小时，压缩补偿会增大其幅度；当信号较大时，则会减小其幅度。

首先测量 pn 二极管结的电压，然后通过电路校正可怕的温度漂移。大多数压缩扩展电路采用另一种策略：用一组直线近似非线性关系。典型方案是用16段直线（称为**线段**）近似对数曲线。8位输出的首位指示输入是正还是负，接下来的三位确定使用8个正线段或8个负线段中的哪一个，最后四位将每个线段划分为16个等距增量。与大多数集成电路一样，压缩扩展芯片拥有复杂且专有的内部设计。与其纠结芯片内部构造，不如重点关注引脚布局和规格说明书。

语音合成与识别

语音生成与识别是计算机领域极具挑战性的难题，尽管尝试过多种方法，但收效甚微。作为数字信号处理（DSP）研究的前沿领域，这一课题未来多年仍将保持重要地位。若指望本节详细讲解语音合成与识别电路的构建方法，您可能会大失所望。这里仅作简要介绍典型方法。需要说明的是，市面上大多数能模拟人声的商用产品并非真正合成语音，而是直接播放人类说话者录制的数字音频片段。这种方法虽然音质出色，但受限于预录词汇和短语的使用范围。

几乎所有语音合成与识别技术都基于图22-8所示的人类语音生成模型。人类语音大多可分为**浊音**和**摩擦音**两类。浊音产生时，空气从肺部被挤压，经声带后从口腔或鼻腔排出。声带是由两片薄薄的组织瓣构成，横跨在

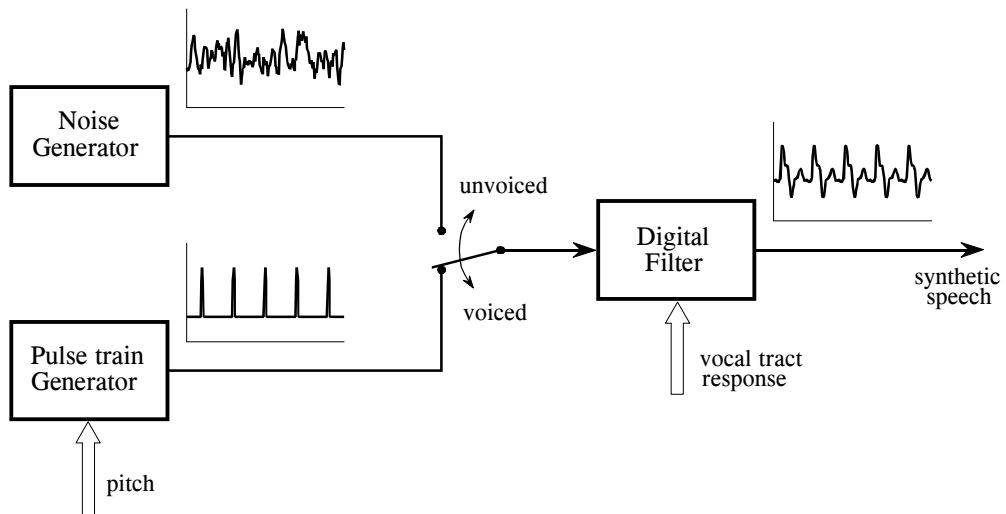


FIGURE 22-8

Human speech model. Over a short segment of time, about 2 to 40 milliseconds, speech can be modeled by three parameters: (1) the selection of either a periodic or a noise excitation, (2) the pitch of the periodic excitation, and (3) the coefficients of a recursive linear filter mimicking the vocal tract response.

the air flow, just behind the Adam's apple. In response to varying muscle tension, the vocal cords vibrate at frequencies between 50 and 1000 Hz, resulting in periodic puffs of air being injected into the throat. Vowels are an example of voiced sounds. In Fig. 22-8, voiced sounds are represented by the pulse train generator, with the pitch (i.e., the fundamental frequency of the waveform) being an adjustable parameter.

In comparison, *fricative* sounds originate as random noise, not from vibration of the vocal cords. This occurs when the air flow is nearly blocked by the tongue, lips, and/or teeth, resulting in air turbulence near the constriction. Fricative sounds include: *s*, *f*, *sh*, *z*, *v*, and *th*. In the model of Fig. 22-8, fricatives are represented by a *noise generator*.

Both these sound sources are modified by the acoustic cavities formed from the tongue, lips, mouth, throat, and nasal passages. Since sound propagation through these structures is a linear process, it can be represented as a linear filter with an appropriately chosen impulse response. In most cases, a *recursive* filter is used in the model, with the recursion coefficients specifying the filter's characteristics. Because the acoustic cavities have dimensions of several centimeters, the frequency response is primarily a series of resonances in the kilohertz range. In the jargon of audio processing, these resonance peaks are called the **formant frequencies**. By changing the relative position of the tongue and lips, the formant frequencies can be changed in both frequency and amplitude.

Figure 22-9 shows a common way to display speech signals, the **voice spectrogram**, or **voiceprint**. The audio signal is broken into short segments,

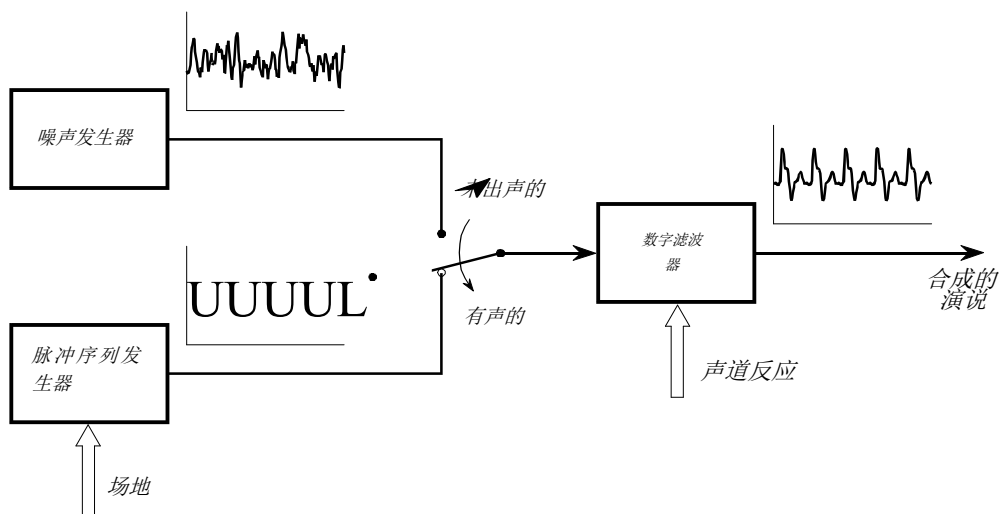


图22-8

人类语音模型。在短时间内（约2至40毫秒），语音可通过三个参数进行建模：(1) 选择周期性激励或噪声激励，(2) 周期性激励的音高，以及 (3) 模拟声道响应的递归线性滤波器系数。

气流直接作用于喉结后方。随着肌肉张力的变化，声带以50至1000赫兹的频率振动，从而产生周期性向喉部喷射气流的脉冲。元音是浊音的典型代表。如图22-8所示，脉冲序列发生器可模拟浊音，其中音高（即波形的基本频率）为可调节参数。

相比之下，摩擦音源于随机噪声，而非声带振动。当气流被舌头、嘴唇和/或牙齿几乎阻塞时，会在狭窄处产生空气湍流，从而产生摩擦音。摩擦音包括：*s*、*f*、*sh*、*z*、*v*和*th*。在图22-8的模型中，摩擦音由噪声发生器表示。

这两种声源都会受到舌头、嘴唇、口腔、咽喉及鼻腔等声学腔体的修饰。由于声音在这些结构中的传播是线性过程，因此可以用具有适当冲激响应的线性滤波器来表征。在大多数情况下，模型中采用的是递归滤波器，其递归系数决定了滤波器的特性。由于声学腔体的尺寸可达数厘米，其频率响应主要表现为千赫兹范围内的共振峰。在音频处理领域，这些共振峰被称为**格式频率**。通过调整舌头与嘴唇的相对位置，可以改变格式频率的频率和振幅。

图22-9展示了一种常见的语音信号显示方式，即**语音频谱图**或**声纹**。音频信号被分割成短片段，

say 2 to 40 milliseconds, and the FFT used to find the frequency spectrum of each segment. These spectra are placed side-by-side, and converted into a grayscale image (low amplitude becomes light, and high amplitude becomes dark). This provides a graphical way of observing how the frequency content of speech changes with time. The segment length is chosen as a tradeoff between *frequency resolution* (favored by longer segments) and *time resolution* (favored by shorter segments).

As demonstrated by the *a* in *rain*, voiced sounds have a periodic time domain waveform, shown in (a), and a frequency spectrum that is a series of regularly spaced harmonics, shown in (b). In comparison, the *s* in *storm*, shows that fricatives have a noisy time domain signal, as in (c), and a noisy spectrum, displayed in (d). These spectra also show the shaping by the formant frequencies for both sounds. Also notice that the time-frequency display of the word *rain* looks similar both times it is spoken.

Over a short period, say 25 milliseconds, a speech signal can be approximated by specifying three parameters: (1) the selection of either a periodic or random noise excitation, (2) the frequency of the periodic wave (if used), and (3) the coefficients of the digital filter used to mimic the vocal tract response. Continuous speech can then be synthesized by continually updating these three parameters about 40 times a second. This approach was responsible for one the early commercial successes of DSP: the *Speak & Spell*, a widely marketed electronic learning aid for children. The sound quality of this type of speech synthesis is poor, sounding very mechanical and not quite human. However, it requires a very low data rate, typically only a few kbits/sec.

This is also the basis for the **linear predictive coding (LPC)** method of speech compression. Digitally recorded human speech is broken into short segments, and each is characterized according to the three parameters of the model. This typically requires about a dozen bytes per segment, or 2 to 6 kbytes/sec. The segment information is transmitted or stored as needed, and then reconstructed with the speech synthesizer.

Speech recognition algorithms take this a step further by trying to recognize patterns in the extracted parameters. This typically involves comparing the segment information with templates of previously stored sounds, in an attempt to identify the spoken words. The problem is, this method does not work very well. It is useful for some applications, but is far below the capabilities of human listeners. To understand why speech recognition is so difficult for computers, imagine someone unexpectedly speaking the following sentence:

Larger run medical buy dogs fortunate almost when.

Of course, you will not understand the meaning of this sentence, because it has none. More important, you will probably not even understand all of the individual words that were spoken. This is basic to the way that humans

将语音片段的时长设定为2至40毫秒，并使用FFT获取每个片段的频谱。这些频谱并排排列后转换为灰度图像（低振幅对应亮色，高振幅对应暗色）。这种可视化方式直观呈现了语音频率成分随时间变化的规律。片段长度的选择需要权衡**频率分辨率**（较长片段更优）与**时间分辨率**（较短片段更优）之间的关系。

如*a*在*rain*中的发音所示，浊音在时域呈现周期性波形（图a），其频谱由一系列等间距谐波构成（图b）。相比之下，*s*在*storm*中的摩擦音则表现为时域信号噪声较大（图c），频谱同样存在明显杂波（图d）。这些频谱还显示出两种发音的格式频率对音形的塑造作用。值得注意的是，单词*rain*的时频显示在两种发音方式下呈现相似特征。

在短短25毫秒内，语音信号可通过设定三个参数进行近似模拟：(1) 选择周期性或随机噪声激励源，(2) 周期性波的频率（若使用），(3) 用于模拟声道响应的数字滤波器系数。通过每秒约40次持续更新这三个参数，即可实现连续语音的合成。这种技术曾助力数字信号处理（DSP）早期商业成功案例——广受儿童市场欢迎的电子学习设备*Speak&Spell*。此类语音合成音质欠佳，听起来机械感十足且不够自然。不过其数据传输速率极低，通常仅需几kb/s。

这正是语音压缩技术中**线性预测编码**（LPC）方法的理论基础。具体来说，数字录制的人声会被分割成多个短片段，每个片段都会根据模型的三个参数进行特征提取。通常每个片段需要约12字节的数据量，即每秒2到6千字节。这些片段信息会根据实际需求进行传输或存储，之后再由语音合成器进行重建。

语音识别算法更进一步，试图从提取的参数中识别模式。这通常需要将语音片段信息与预先存储的声音模板进行比对，以识别出实际说出的词语。但问题在于，这种方法效果并不理想。虽然在某些应用场景中仍有一定用武之地，但其识别能力远不及人类听觉。要理解计算机为何难以实现语音识别，不妨想象有人突然说出以下句子：

大型犬只医疗采购项目在幸运时几乎成功。

当然，你无法理解这句话的含义，因为它本身毫无意义。更重要的是，你甚至可能听不清其中任何一个单词。这正是人类认知方式的基本特征。

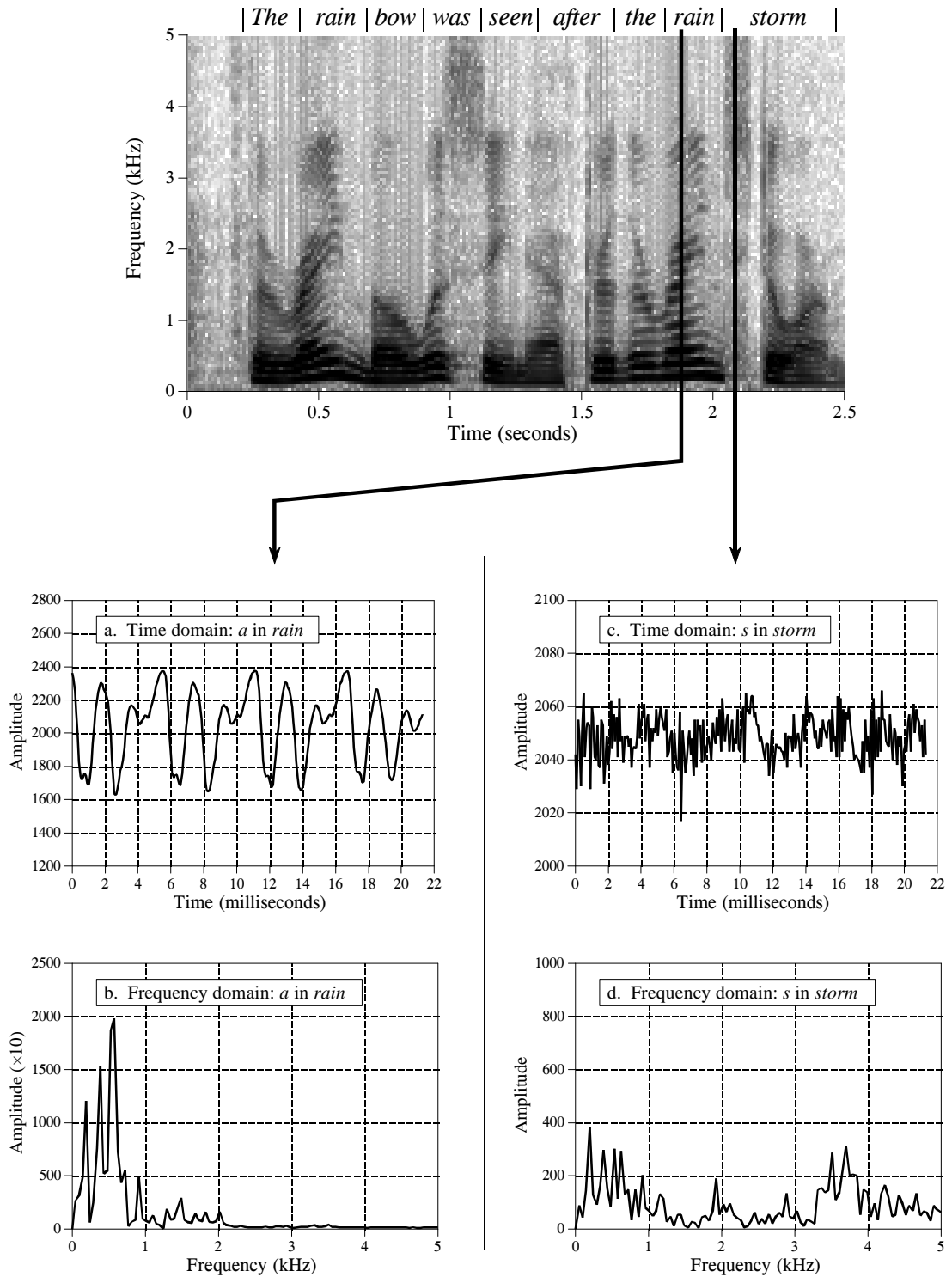


FIGURE 22-9

Voice spectrogram. The spectrogram of the phrase: "The rainbow was seen after the rain storm." Figures (a) and (b) shows the time and frequency signals for the voiced *a* in *rain*. Figures (c) and (d) show the time and frequency signals for the fricative *s* in *storm*.

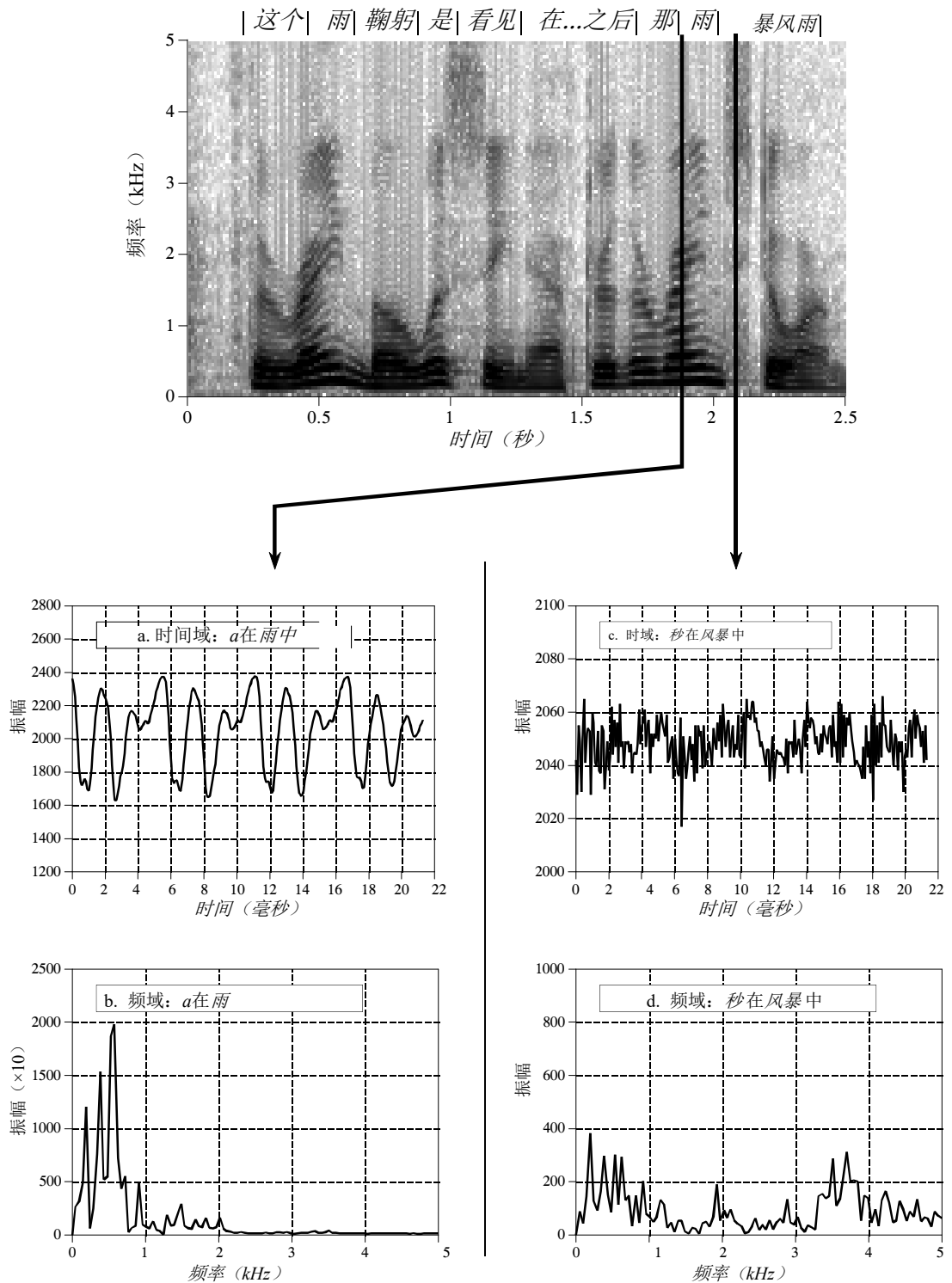


图22-9

语音频谱图。短语“雨后天见彩虹”的频谱图。图(a)和(b)显示了rain中浊音*a*的时频信号。图(c)和(d)显示了storm中摩擦音*s*的时频信号。

perceive and understand speech. Words are recognized by their sounds, but also by the *context* of the sentence, and the *expectations* of the listener. For example, imagine hearing the two sentences:

The child wore a spider ring on Halloween.

He was an American spy during the war.

Even if exactly the same sounds were produced to convey the underlined words, listeners *hear* the correct words for the context. From your accumulated knowledge about the world, you know that children don't wear secret agents, and people don't become spooky jewelry during wartime. This usually isn't a conscious act, but an inherent part of human hearing.

Most speech recognition algorithms rely only on the sound of the individual words, and not on their context. They attempt to *recognize words*, but not to *understand speech*. This places them at a tremendous disadvantage compared to human listeners. Three annoyances are common in speech recognition systems: (1) The recognized speech must have distinct pauses between the words. This eliminates the need for the algorithm to deal with phrases that sound alike, but are composed of different words (i.e., *spider ring* and *spy during*). This is slow and awkward for people accustomed to speaking in an overlapping flow. (2) The vocabulary is often limited to only a few hundred words. This means that the algorithm only has to search a limited set to find the best match. As the vocabulary is made larger, the recognition time and error rate both increase. (3) The algorithm must be *trained* on each speaker. This requires each person using the system to speak each word to be recognized, often needing to be repeated five to ten times. This personalized database greatly increases the accuracy of the word recognition, but it is inconvenient and time consuming.

The prize for developing a successful speech recognition technology is enormous. Speech is the quickest and most efficient way for humans to communicate. Speech recognition has the potential of replacing writing, typing, keyboard entry, and the electronic control provided by switches and knobs. It just needs to work a little better to become accepted by the commercial marketplace. Progress in speech recognition will likely come from the areas of artificial intelligence and neural networks as much as through DSP itself. Don't think of this as a technical *difficulty*; think of it as a technical *opportunity*.

Nonlinear Audio Processing

Digital filtering can improve audio signals in many ways. For instance, *Wiener filtering* can be used to separate frequencies that are mainly signal, from frequencies that are mainly noise (see Chapter 17). Likewise, *deconvolution* can compensate for an undesired convolution, such as in the restoration of old

感知和理解语言。单词可以通过它们的声音来识别，也可以通过句子的上下文和听者的期望来识别。例如，想象一下听到以下两个句子：

这个孩子在万圣节那天戴着蜘蛛戒指。

他是战争期间的美国间谍。

即使为了传达划线单词而发出完全相同的声音，听者也能听出正确的单词。根据你对世界的积累，你知道孩子们不会穿特工的衣服，人们也不会战争时期变成吓人的珠宝。这通常不是一种有意识的行为，而是人类听觉的固有部分。

大多数语音识别算法仅依赖单词的发音特征，而忽视语境信息。它们试图识别词汇，却无法理解语音内容，这使得它们在与人类听者对比时处于巨大劣势。语音识别系统存在三大痛点：(1)语音识别必须区分单词间的停顿。这虽然避免了算法处理发音相似但由不同词汇组成的短语（如*spider ring*和*spy during*），但对于习惯连续说话的人而言，这种处理方式既缓慢又生硬。(2)词汇库通常仅包含数百个单词。这意味着算法只需在有限词汇中寻找最佳匹配，但随着词汇量增加，识别时间和错误率都会攀升。(3)算法需要针对每个说话者进行训练。这意味着每位使用者的每个发音都需要重复五到十次才能被识别。这种个性化数据库极大地提高了单词识别的准确性，但其操作不便且耗时。

开发成功语音识别技术的回报是巨大的。语音是人类最快捷、最高效的交流方式。语音识别技术有望取代书写、打字、键盘输入以及开关旋钮提供的电子控制。它只需进一步完善就能被商业市场接受。语音识别的进步很可能既来自数字信号处理技术本身，也来自人工智能和神经网络领域。不要将此视为技术难题，而应视为技术机遇。

非线性音频处理

数字滤波可以通过多种方式改善音频信号。例如，*Wiener*滤波可以用来分离主要为信号的频率和主要为噪声的频率（参见第17章）。同样，去卷积可以补偿不希望的卷积，例如在恢复旧的

recordings (also discussed in Chapter 17). These types of linear techniques are the backbone of DSP. Several *nonlinear* techniques are also useful for audio processing. Two will be briefly described here.

The first nonlinear technique is used for reducing wideband noise in speech signals. This type of noise includes: magnetic tape hiss, electronic noise in analog circuits, wind blowing by microphones, cheering crowds, etc. Linear filtering is of little use, because the frequencies in the noise completely overlap the frequencies in the voice signal, both covering the range from 200 hertz to 3.2 kHz. How can two signals be separated when they overlap in both the time domain *and* the frequency domain?

Here's how it is done. In a short segment of speech, the amplitude of the frequency components are greatly *unequal*. As an example, Fig. 22-10a illustrates the frequency spectrum of a 16 millisecond segment of speech (i.e., 128 samples at an 8 kHz sampling rate). Most of the signal is contained in a few large amplitude frequencies. In contrast, (b) illustrates the spectrum when only random noise is present; it is very irregular, but more uniformly distributed at a low amplitude.

Now the key concept: if both signal and noise are present, the two can be partially separated by looking at the *amplitude* of each frequency. If the amplitude is large, it is probably mostly signal, and should therefore be retained. If the amplitude is small, it can be attributed to mostly noise, and should therefore be discarded, i.e., set to zero. Mid-size frequency components are adjusted in some smooth manner between the two extremes.

Another way to view this technique is as a *time varying Wiener filter*. As you recall, the frequency response of the Wiener filter passes frequencies that are mostly signal, and rejects frequencies that are mostly noise. This

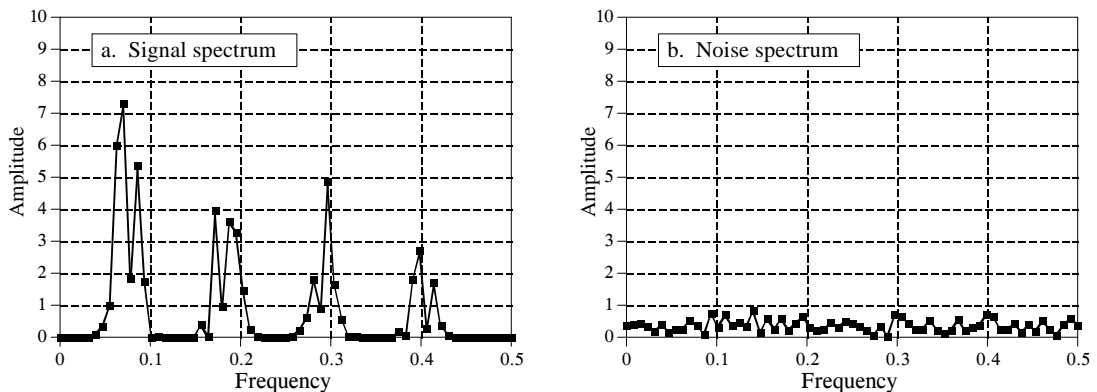


FIGURE 22-10

Spectra of speech and noise. While the frequency spectra of speech and noise generally overlap, there is some separation if the signal segment is made short enough. Figure (a) illustrates the spectrum of a 16 millisecond speech segment, showing that many frequencies carry little speech information, *in this particular segment*. Figure (b) illustrates the spectrum of a random noise source; all the components have a small amplitude. (These graphs are not of real signals, but illustrations to show the noise reduction technique).

录音（详见第17章）。这类线性技术是数字信号处理（DSP）的核心。若干非线性技术同样适用于音频处理，本文将简要介绍其中两种。

第一种非线性技术用于降低语音信号中的宽带噪声。这类噪声包括：磁带静电声、模拟电路中的电子噪声、麦克风前的风声、人群欢呼声等。线性滤波在此类情况下效果有限，因为噪声频率与语音信号频率完全重叠，两者都覆盖200赫兹至3.2千赫兹的频段。当两个信号在时域和频域都存在重叠时，如何实现有效分离？

具体实现方式如下：在一段短语音片段中，各频率分量的振幅呈现显著不平衡状态。例如图22-10a展示了16毫秒语音片段（即8kHz采样率下128个样本）的频谱分布，大部分信号集中在少数高频段。相比之下，图(b)显示了仅存在随机噪声时的频谱特征——虽然整体分布极不规则，但在低振幅区域呈现出更为均匀的分布模式。

现在说说关键概念：当信号和噪声同时存在时，可以通过分析各频率的振幅进行部分分离。若振幅较大，说明信号占主导，应保留；若振幅较小，则可能主要为噪声，应舍弃（即设为零）。中等振幅的频率成分则会在两者之间以平滑方式调整。

另一种理解该技术的方式是将其视为时变维纳滤波器。正如你所记得的，维纳滤波器的频率响应会通过大部分信号频率，同时抑制大部分噪声频率。

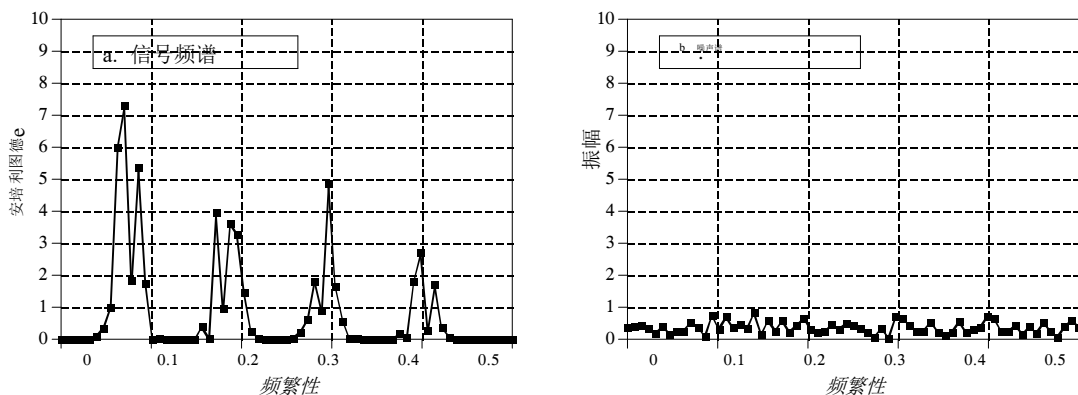


图22-10

语音与噪声的频谱分析。虽然语音和噪声的频谱通常存在重叠，但若将信号片段足够短，两者之间仍会形成一定分离。图(a)展示了16毫秒语音片段的频谱，在该特定片段中可见许多频率承载的语音信息量极少。图(b)则呈现了随机噪声源的频谱特征，所有成分的振幅均较小。（需说明这些图表并非真实信号，仅用于说明降噪技术原理的示意图）。

requires a knowledge of the signal and noise spectra *beforehand*, so that the filter's frequency response can be determined. This nonlinear technique uses the same idea, except that the Wiener filter's frequency response is recalculated for each segment, based on the spectrum *of that segment*. In other words, the filter's frequency response changes from segment-to-segment, as determined by the characteristics of the signal itself.

One of the difficulties in implementing this (and other) nonlinear techniques is that the overlap-add method for filtering long signals is not valid. Since the frequency response changes, the time domain waveform of each segment will no longer align with the neighboring segments. This can be overcome by remembering that audio information is encoded in frequency patterns that change over time, and not in the shape of the time domain waveform. A typical approach is to divide the original time domain signal into *overlapping* segments. After processing, a smooth window is applied to each of the overlapping segments before they are recombined. This provides a smooth transition of the frequency spectrum from one segment to the next.

The second nonlinear technique is called **homomorphic** signal processing. This term literally means: *the same structure*. Addition is not the only way that noise and interference can be combined with a signal of interest; multiplication and convolution are also common means of mixing signals together. If signals are combined in a nonlinear way (i.e., anything other than addition), they cannot be separated by linear filtering. Homomorphic techniques attempt to separate signals combined in a nonlinear way by making the problem *become* linear. That is, the problem is converted to the *same structure* as a linear system.

For example, consider an audio signal transmitted via an AM radio wave. As atmospheric conditions change, the received amplitude of the signal increases and decreases, resulting in the loudness of the received audio signal slowly changing over time. This can be modeled as the audio signal, represented by $a[n]$, being *multiplied* by a slowly varying signal, $g[n]$, that represents the changing gain. This problem is usually handled in an electronic circuit called an *automatic gain control* (AGC), but it can also be corrected with nonlinear DSP.

As shown in Fig. 22-11, the input signal, $a[n] \times g[n]$, is passed through the logarithm function. From the identity, $\log(x \times y) = \log x + \log y$, this results in two signals that are combined by addition, i.e., $\log a[n] + \log g[n]$. In other words, the *logarithm* is the homomorphic transform that turns the nonlinear problem of *multiplication* into the linear problem of *addition*.

Next, the added signals are separated by a conventional linear filter, that is, some frequencies are passed, while others are rejected. For the AGC, the gain signal, $g[n]$, will be composed of very low frequencies, far below the 200 hertz to 3.2 kHz band of the voice signal. The logarithm of these signals will have more complicated spectra, but the idea is the same: a high-pass filter is used to eliminate the varying gain component from the signal.

该方法需要预先掌握信号与噪声的频谱特征，从而确定滤波器的频率响应。这种非线性技术采用相同原理，但其创新之处在于：针对每个信号片段，都会根据其频谱特征重新计算维纳滤波器的频率响应。换言之，滤波器的频率响应会随着信号本身的特性变化而逐段调整。

在实施这种（及其他）非线性技术时，一个主要难点在于滤波长信号的重叠相加法并不适用。由于频率响应会发生变化，各段信号的时间域波形将不再与相邻段对齐。解决这一问题的关键在于：音频信息是以随时间变化的频率模式编码的，而非基于时间域波形的形态。典型处理方法是将原始时域信号划分为重叠的多个段落。经过处理后，对每个重叠段应用平滑窗口处理，再进行重新组合。这种处理方式能确保频谱在相邻段落间实现平滑过渡。

第二种非线性技术被称为**同态信号处理**。这个术语的字面意思是“相同结构”。噪声和干扰与目标信号结合的方式不仅限于加法运算，乘法和卷积也是常见的信号混合手段。当信号以非线性方式（即除加法外的任何方式）组合时，它们无法通过线性滤波进行分离。同态技术试图通过将问题转化为线性问题来分离非线性组合的信号，即通过将问题转换为与线性系统具有相同结构的形式。

例如，考虑通过调幅无线电波传输的音频信号。当大气条件变化时，接收到的信号幅度会随之增减，导致接收音频信号的响度随时间缓慢变化。这可以建模为：由 $a[n]$ 表示的音频信号，被乘以一个缓慢变化的信号 $g[n]$ ，该信号代表增益的变化。这类问题通常通过称为**自动增益控制**（AGC）的电子电路处理，但也可通过非线性数字信号处理（DSP）进行校正。

如图22-11所示，输入信号 $a[n] \times g[n]$ 经过对数函数处理。根据恒等式 $\log(x+y) = \log x + \log y$ ，由此得到两个通过加法组合的信号，即 $\log a[n] + \log g[n]$ 。换言之，对数是将乘法的非线性问题转化为加法的线性问题的同态变换。

随后，这些新增信号通过常规线性滤波器进行分离，即部分频率被保留，而其他频率则被抑制。对于AGC而言，增益信号 $g[n]$ 将包含远低于语音信号200赫兹至3.2千赫频段的极低频成分。这些信号的对数将呈现更复杂的频谱，但原理相同：采用高通滤波器来消除信号中的动态增益分量。

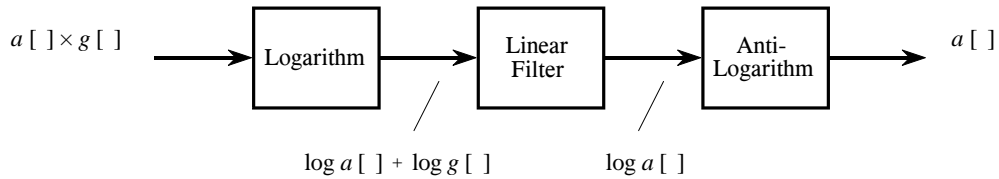


FIGURE 22-11

Homomorphic separation of multiplied signals. Taking the logarithm of the input signal transforms components that are *multiplied* into components that are *added*. These components can then be separated by linear filtering, and the effect of the logarithm undone.

In effect, $\log a[n] + \log g[n]$ is converted into $\log a[n]$. In the last step, the logarithm is undone by using the exponential function (the anti-logarithm, or e^x), producing the desired output signal, $a[n]$.

Figure 22-12 shows a homomorphic system for separating signals that have been *convolved*. An application where this has proven useful is in removing echoes from audio signals. That is, the audio signal is convolved with an impulse response consisting of a delta function plus a shifted and scaled delta function. The homomorphic transform for convolution is composed of two stages, the *Fourier transform*, changing the convolution into a multiplication, followed by the *logarithm*, turning the multiplication into an addition. As before, the signals are then separated by linear filtering, and the homomorphic transform undone.

An interesting twist in Fig. 22-12 is that the linear filtering is dealing with frequency domain signals in the same way that time domain signals are usually processed. In other words, the time and frequency domains have been swapped from their normal use. For example, if FFT convolution were used to carry out the linear filtering stage, the "spectra" being multiplied would be in the *time domain*. This role reversal has given birth to a strange jargon. For instance, *cepstrum* (a rearrangement of *spectrum*) is the Fourier transform of the logarithm of the Fourier transform. Likewise, there are *long-pass* and *short-pass* filters, rather than low-pass and high-pass filters. Some authors even use *Quefrency Analysis* and *liftering*.

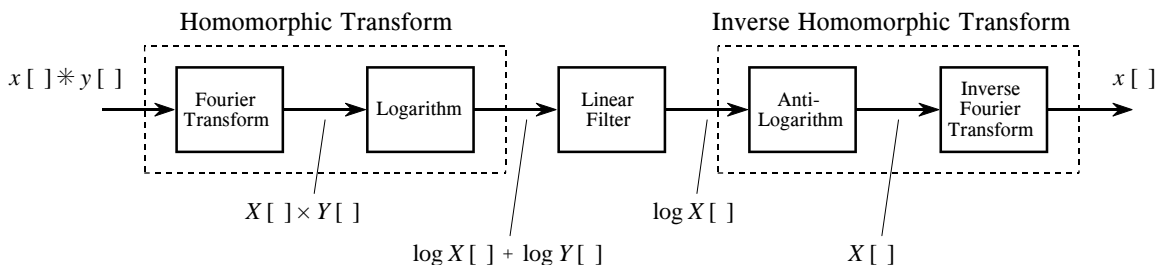


FIGURE 22-12

Homomorphic separation of convolved signals. Components that have been *convolved* are converted into components that are *added* by taking the Fourier transform followed by the logarithm. After linear filtering to separate the added components, the original steps are undone.

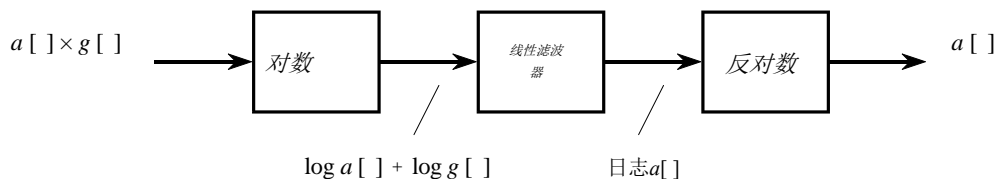


图22-11

乘积信号的同态分离。对输入信号取对数，将相乘的分量转换为相加的分量。这些分量可以通过线性滤波分离，从而消除对数的影响。

实际上， $\log a[] + \log g[]$ 被转换为 $\log a[]$ 。在最后一步，通过使用指数函数（反对数，或 e^x ）对数进行逆运算，生成所需的输出信号， $a[]$ 。

图22-12展示了一个用于分离经过卷积处理信号的同态变换系统。该技术在音频信号回声消除领域展现出显著优势：具体而言，音频信号会与由 δ 函数叠加移位缩放后的 δ 函数构成的冲激响应进行卷积运算。其同态变换过程包含两个关键步骤：首先通过傅里叶变换将卷积转化为乘法运算，随后借助对数变换将乘法运算转换为加法运算。最终通过线性滤波实现信号分离，并对同态变换进行逆向处理。

图22-12中一个有趣的转折在于，线性滤波处理频域信号的方式与时域信号通常处理方式相同。换言之，时域和频域已从其常规用途中被调换。例如，若使用FFT卷积来执行线性滤波阶段，被相乘的“频谱”将处于时域。这种角色反转催生了一套奇特的术语体系。例如倒谱（频谱的重新排列）是傅里叶变换的对数的傅里叶变换。同样存在长通和短通滤波器，而非传统的低通和高通滤波器。部分学者甚至使用频率分析和拉伸等术语。

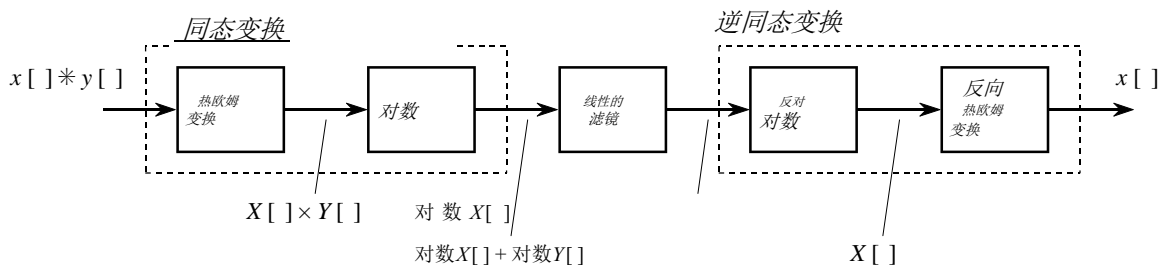


图22-12

卷积信号的同态分离。卷积的分量通过傅里叶变换后取对数加上，转换为分量。经过线性滤波分离加上的分量后，恢复原始步骤。

Keep in mind that these are simplified descriptions of sophisticated DSP algorithms; homomorphic processing is filled with subtle details. For example, the logarithm must be able to handle both negative and positive values in the input signal, since this is a characteristic of audio signals. This requires the use of the *complex logarithm*, a more advanced concept than the logarithm used in everyday science and engineering. When the linear filtering is restricted to be a *zero phase* filter, the complex log is found by taking the simple logarithm of the absolute value of the signal. After passing through the zero phase filter, the sign of the original signal is reapplied to the filtered signal.

Another problem is *aliasing* that occurs when the logarithm is taken. For example, imagine digitizing a continuous *sine wave*. In accordance with the sampling theorem, two or more samples per cycle is sufficient. Now consider digitizing the logarithm of this continuous sine wave. The sharp corners require many more samples per cycle to capture the waveform, i.e., to prevent aliasing. The required sampling rate can easily be 100 times as great after the log, as before. Further, it doesn't matter if the logarithm is applied to the continuous signal, or to its digital representation; the result is the same. Aliasing will result unless the sampling rate is high enough to capture the sharp corners produced by the nonlinearity. The result is that audio signals may need to be sampled at 100 kHz or more, instead of only the standard 8 kHz.

Even if these details are handled, there is no guarantee that the linearized signals *can* be separated by the linear filter. This is because the spectra of the linearized signals can overlap, even if the spectra of the original signals do not. For instance, imagine adding two sine waves, one at 1 kHz, and one at 2 kHz. Since these signals do not overlap in the frequency domain, they can be completely separated by linear filtering. Now imagine that these two sine waves are multiplied. Using homomorphic processing, the log is taken of the combined signal, resulting in the log of one sine wave plus the log of the other sine wave. The problem is, the logarithm of a sine wave contains many harmonics. Since the harmonics from the two signals overlap, their complete separation is not possible.

In spite of these obstacles, homomorphic processing teaches an important lesson: signals should be processed in a manner *consistent* with how they are formed. Put another way, the first step in any DSP task is to understand how information is represented in the signals being processed.

需要特别说明的是，这些描述只是对复杂数字信号处理算法的简化说明，同态处理技术中蕴含着诸多精妙细节。例如，由于音频信号的特性，对数运算必须能够处理输入信号中的正负值。这就需要运用复对数——这个概念比日常科学工程中使用的对数更为复杂。当线性滤波器被限定为零相位滤波器时，复对数的计算方式就是对信号绝对值取简单对数。经过零相位滤波器处理后，原始信号的正负符号会被重新赋予到滤波后的信号上。

另一个问题是取对数时出现的混叠现象。例如，假设对连续的正弦波进行数字化处理。根据采样定理，每周期取两个或更多样本就足够了。但若对这个连续正弦波的对数进行数字化，由于存在尖锐的拐角，就需要每周采集更多样本才能完整捕捉波形，即防止混叠。此时所需的采样率可能比原始信号高出100倍。此外，无论对连续信号还是其数字表示取对数，结果都相同。只有当采样率足够高以捕捉非线性产生的尖锐拐角时，混叠才会发生。这意味着音频信号可能需要以100kHz甚至更高的采样率进行处理，而不仅仅是标准的8kHz。

即使处理了这些细节，也不能保证线性化信号能通过线性滤波器实现分离。这是因为线性化信号的频谱可能会重叠，即便原始信号的频谱没有重叠。举个例子，假设叠加两个正弦波，一个频率为1千赫兹，另一个为2千赫兹。由于这两个信号在频域中不重叠，通过线性滤波就能完全分离。但若将这两个正弦波相乘，使用同态处理时对合成信号取对数，就会得到一个正弦波的对数加上另一个正弦波的对数。问题在于，正弦波的对数包含大量谐波成分。由于两个信号的谐波存在重叠，因此无法实现完全分离。

尽管存在这些障碍，同态处理给我们上了一堂重要的课：信号的处理方式应该与它们的形成方式一致。换句话说，任何DSP任务的第一步都是理解信息是如何在被处理的信号中表示的。